Toward Optimal Deployment of Cloud-Assisted Video Distribution Services

Jian He, Di Wu, *Member, IEEE,* Yupeng Zeng, Xiaojun Hei, *Member, IEEE,* and Yonggang Wen

***Abstract*—For Internet video services, the high fluctuation of user demands in geographically distributed regions results in low resource utilizations of traditional content distribution network systems. Due to the capability of rapid and elastic resource provisioning, cloud computing emerges as a new paradigm to reshape the model of video distribution over the Internet, in which resources (such as bandwidth, storage) can be rented on demand from cloud data centers to meet volatile user demands. However, it is challenging for a video service provider (VSP) to optimally deploy its distribution infrastructure over multiple geo-distributed cloud data centers. A VSP needs to minimize the operational cost induced by the rentals of cloud resources without sacrificing user experience in all regions. The geographical diversity of cloud resource prices further makes the problem complicated. In this paper, we investigate the optimal deployment problem of cloud-assisted video distribution services and explore the best tradeoff between the operational cost and the user experience. We aim to pave the way for building the next-generation video cloud. Toward this objective, we first formulate the deployment problem into a min-cost network flow problem, which takes both the operational cost and the user experience into account. Then, we apply the Nash bargaining solution to solve the joint optimization problem efficiently and derive the optimal bandwidth provisioning strategy and optimal video placement strategy. In addition, we extend the algorithms to the online case and consider the scenario when peers participate into video distribution. Finally, we conduct extensive simulations to evaluate our algorithms in the realistic settings. Our results show that our proposed algorithms can achieve a good balance among multiple objectives and effectively optimize both operational cost and user experience.**

***Index Terms*—Cloud deployment, Nash bargaining solution, video distribution.**

Manuscript received October 5, 2012; revised December 19, 2012 and February 2, 2013; accepted March 14, 2013. Date of publication March 28, 2013; date of current version September 28, 2013. This work was supported in part by the NSFC, under Grants 60972014, 61003242, 61272397, the Fundamental Research Funds for the Central Universities, under Grants HUST:2012TS018, 12LGPY53, the National Technology Support Plan of China, under Grant 2011BAK08B00, the Guangdong Natural Science Funds for Distinguished Young Scholar, under Grant S20120011187, the Program for New Century Excellent Talents in University, under Grant NCET-11-0542, and the Guangzhou Pearl River Science and Technology Rising Star Project, under Grant 2011J2200086. This paper was recommended by Associate Editor W. Zeng. *(Corresponding author: D. Wu.)*

J. He, D. Wu, and Y. Zeng are with the Department of Computer Science, Sun Yat-Sen University, Guangzhou 510006, China (e-mail: hejian9@mail2.sysu.edu.cn; wudi27@mail.sysu.edu.cn; zengyp@mail2.sysu.edu.cn).

X. Hei is with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: heixj@hust.edu.cn).

Y. Wen is with the School of Computer Engineering, Nanyang Technological University, Singapore (e-mail: ygwen@ntu.edu.sg).

Digital Object Identifier 10.1109/TCSVT.2013.2255423

I. Introduction

In the past decade, Internet video has become a very popular application over the Internet and accounts for the majority of network traffic [1]. Internet video service providers (e.g., YouTube, Hulu) generally resort to a content distribution network (CDN) to conduct large-scale video distribution. However, CDN solutions are inadequate for the emerging video traffic growth. First, due to their semistatic resource provisioning mechanism, the resource utilization of existing CDNs is extremely low (normally ranging between 5% and 10%) [2], which directly translates into a high operational cost. Especially, significant oversubscription has been observed for flash-crowd situations [3] when a lot of users are trying to consume the same video content. Second, the emerging user-generated video contents (e.g., Youtube, Ku6, etc.) are long-tail nature [4], defying the operational principle of serving the popular contents by CDNs. Therefore, new paradigms should be developed in order to provide the capabilities of scaling up and down the provisioned resources in a dynamic manner and improve the resource utilization ratio.

The emergence of cloud computing [5] opens a new door for designing the next-generation video distribution platform. Cloud-based services are more cost-effective, highly scalable, and reliable. Recently, there have been quite a few research works (e.g., [6]–[13]) on exploiting the cloud platform for media content distribution. Compared with conventional approaches, a unique feature of cloud-assisted video distribution is the capability of rapid and elastic resource provisioning. Cloud resources, such as CPU, memory, storage, and bandwidth, can be automatically allocated in a fine granularity to meet the demand from end users timely. A video service provider can rent the distribution infrastructure from the cloud service provider (CSP) in an on-demand manner, and avoid the long-term hardware investment and low utilization due to resource over-provisioning.

However, for bandwidth-intensive services such as video distribution, a number of challenging issues need to be addressed before real deployment on cloud platforms. First, as users are spreading over multiple regions, to improve user quality of experience at different regions, a video service provider should deploy its distribution service in multiple geographically distributed data centers, which are possibly owned by multiple CSPs with different pricing strategies. Thus, for a video service provider (VSP), a set of critical questions need to be clearly answered: how should a video

service provider minimize its operational cost? How much resources (e.g., bandwidth) should be provisioned at each location? How should user requests be directed to different geo-distributed data centers so as to maximize overall user experience? Second, the popularities of videos to be distributed are dynamic and evolutionary over time. Thus, the deployment of cloud resources is also a dynamic process. This means that a video service provider should adjust resource provisioning at different regions proactively and place video contents according to the changes of user demands. We need to design online algorithms instead of offline algorithms to handle demand fluctuation.

In this paper, we investigate the optimal deployment of cloud-assisted video distribution services. To this end, we propose a set of practical algorithms to address the above challenges and validate their effectiveness via extensive simulations. Our main contributions in this paper can be summarized as follows.

1) We explicitly formulate the problem of cloud deployment using network flow theory. The optimal deployment problem can be transformed into the problem of finding the min-cost flow in a network flow model. We identify the conflict between optimizing bandwidth provisioning cost and viewing latency and show the related Pareto curve.
2) We investigate the joint optimization problem of cloud deployment and apply the concept of Nash bargaining solution (NBS) to solve it efficiently. The obtained result bears the properties of both optimality and fairness.
3) We design an optimal bandwidth provisioning algorithm with dual decomposition for one-shot optimization, which can be easily extended to the continuous online case. We also derive the associated optimal video content placement strategy.
4) We further study the case when peer bandwidth resources at the edge can be exploited. We consider a locality-aware peer-assisted design and derive the corresponding optimal deployment strategy.
5) We conduct extensive trace-driven simulations to validate the effectiveness of our proposed algorithms in the realistic settings. The experimental results indicate that our algorithms can achieve a good balance among multiple objectives and optimize the operational cost and the user experience simultaneously.

To the best of our knowledge, our work is the first to apply Nash bargaining solution to optimize the bandwidth cost and the latency cost jointly for cloud-assisted video distribution services, which can achieve both optimality and fairness. In addition, our algorithm can be easily extended to incorporate the peer contribution at the edge.

The remainder of this paper is organized as follows. Section II reviews previous work in the area of Internet video distribution and cloud service deployment. Section III presents the formulation of the optimal cloud deployment problem using network flow theory. Section IV considers the optimization of the bandwidth cost and the latency cost jointly when deploying cloud-assisted video services, and proposes an online Nash bargaining algorithm to solve such a joint optimization problem. We also further explore the scenario when peers also participate into video distribution in the local region. Section V proves the effectiveness of our proposed algorithms by experimental evaluation. Finally, Section VI concludes the paper and discusses the future work.

II. Related Work

Internet video distribution has received a great amount of attention in the past decade. CDN and P2P are two widely adopted technologies for building large-scale video distribution platforms. However, CDN is too expensive for small video service providers, while P2P is hard to guarantee viewing quality of end users and becomes inefficient when handling unexpected flash crowd [14]. Xunlei Kankan [15] and LiveSky [16] resort to a hybrid CDN-P2P design to reduce cost without sacrificing user experience. However, there is still a big gap for designing an ideal cost-effective, highly scalable and reliable solution.

The emergence of cloud computing provides a promising approach to deliver QoS-guaranteed multimedia services economically [6]. Virtualization techniques can be exploited for resource allocation in a fine granularity to provide content distribution and multimedia processing service. Initial attempts have been made by academic researchers [7], [8] and industrial practitioners [17], [18] to migrate VoD applications to the cloud. Niu *et al.* [9], [19] further investigated the pricing strategies for VoD service providers and automatic bandwdith allocation to satisfy user demands. In terms of live media streaming, Wang *et al.* [10] presented a generic framework called CALMS for the migration of live streaming services to the cloud, which adaptively leases and adjusts cloud resources to meet dynamic user demands. CloudStream [11] considered realtime transcoding of videos in different qualities over cloud and delivered video streams via a cloud-based SVC proxy. Chen *et al.* [12] investigated the problem of placing Web server replicas in storage clouds to provide cost-effective CDNs. A few recent research work [13], [20]–[22] also investigated the distribution of long-tailed and highly dynamic social media contents over the cloud, and studied how to partition the social contents and conduct resource provisioning more efficiently.

Our work differs from previous work in the following aspects. First, instead of simply defining user latency as a constraint, we focus on the optimization of provisioning cost and user experience jointly, considering that VSPs in different scales have different optimization preferences. In addition, we consider nonlinear cost functions in our problem, which are more generic but incur higher complexity in problem solving. Second, we consider the deployment policy when peer resources can be exploited. The proposed algorithm can be easily extended to meet dynamic user demands. Our work is also different from conventional research work on traffic engineering [23], [24], which focuses on balancing network traffic from the perspective of ISPs. Instead, our algorithms are more tailored to optimize the deployment of cloud servers from the perspective of video service providers.

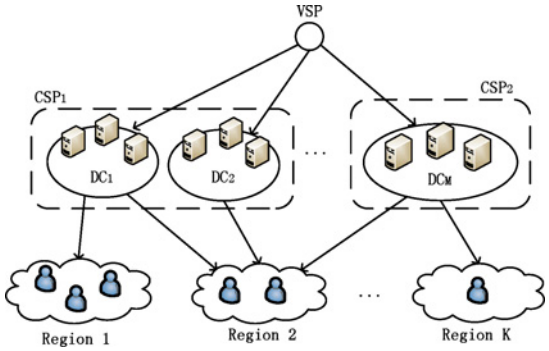


Fig. 1. Simplified cloud-assisted video distribution platform.

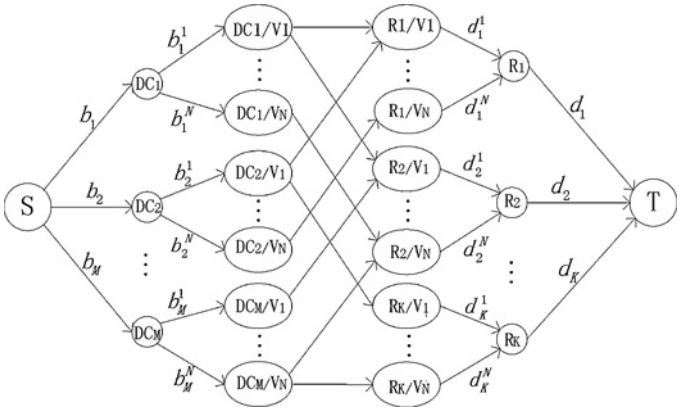


Fig. 2. Network flow model for cloud deployment problem.

III. Problem Formulation

A. Cloud-Assisted Video Distribution Services

In the cloud computing paradigm, a cloud service provider (CSP) can own multiple geographically distributed data centers and manage all the hardware/software resources in each data center for resource pooling. To provide video distribution services, a VSP will rent resources (including CPU, bandwidth, storage) from one or more CSPs, possibly with different resource pricing plans. The allocation of cloud resources is generally in the unit of virtual machine with predefined configurations. By adjusting the number of virtual machines rented at each data center, a VSP can scale up and down the provisioned resources to quickly handle demand fluctuation in different regions. From the perspecive of a VSP, it is challenging to determine how to provision resources at each location dynamically. Each video request generated by end users will be routed to a given data center according to certain routing policies. Such request routing policies should be carefully designed so as to well balance the operational cost and the user experience. For example, if always routing a user request to a local data center, it may lead to a much higher bandwidth cost in spite that such a strategy can provide good user experience.

Let us first consider a typical cloud-assisted video distribution scenario. Suppose there is a VSP that needs to deliver N videos to its users spreading over $K(K \geq 1)$ geographical regions. The video service provider is relying on cloud servers hosted in M data centers to provide video services. Fig. 1 provides an overview of a simplified cloud-assisted video distribution platform.

To optimize the user experience, end users prefer to receive video streams from cloud servers with low latency (e.g., cloud servers deployed in the same region). In case that local bandwidth supply is insufficient, a user can also receive video streams from cloud servers at other regions but at the cost of a higher latency. For a VSP, its objective is to optimize the user experience at all regions and minimize its operational cost at the same time. To this end, a VSP should determine how much resources (e.g., bandwidth, storage, CPU) should be purchased from different CSPs and how to deploy these resources at different regions for better viewing quality.

B. Network Flow Model

Next, we start to study the deployment problem from a simple case and formulate the bandwidth provisioning problem into a min-cost network flow problem. Consider a scenario in which user demand for each video generated from every region is known *a priori*, the bandwidth provisioning problem can be transformed into a network flow model illustrated in Fig. 2. Note that the above assumption is only made for the modeling purpose. In the later section, we will relax the assumption that user demand should be known *a priori* and utilize prediction methods to predict future user demand for practical implementation.

In Fig. 2, node S and T are two virtual nodes that represent the source and destination nodes, which are used to indicate the total bandwidth supply from distributed data centers and the total bandwidth demand generated by all users. The set of data centers involved in video distribution are represented by nodes $\{DC_i, i = 1, \cdots, M\}$. The set of user regions are represented by nodes $\{R_k, k = 1, \cdots, K\}$. To further illustrate the demand and supply for each video in detail, we introduce a set of auxiliary nodes $\{DC_i/V_j, i = 1, \cdots, M, j = 1, \cdots, N\}$ and $\{R_k/V_j, k = 1, \cdots, K, j = 1, \cdots, N\}$. The flow b_i between S and DC_i refers to the cloud bandwidth provisioned at the ith data center, and b_i^j between DC_i and DC_i/V_j refers to the amount of bandwidth allocated to distribute video j among bandwidth b_i, which satisfies $b_i = \sum_{j=1}^{N} b_i^j$. Similarly, d_k denotes the total user demand generated from the kth region and d_k^j between nodes R_k and R_k/V_j denotes the demand of a video j from the kth region and satisfies $d_k = \sum_{j=1}^{N} d_k^j$. The flow b_{ik}^j between nodes DC_i/V_j and R_k/V_j represents the amount of bandwidth supplied by the ith data center to the kth region for distributing video j, which satisfies $b_i = \sum_{j=1}^{N} \sum_{k=1}^{K} b_{ik}^j$.

Each edge in the graph is also associated with a cost. For edges between S and $DC_i(i = 1, \cdots, M)$, the associated cost is mainly bandwidth cost. The unit price of bandwidth can be different for data centers at different locations. According to the pricing strategy of typical CSPs [24], we define the bandwidth cost function $\psi_i(x)$ associated with a data center DC_i as a nondecreasing concave function, where x is the amount of the bandwidth flow supplied by DC_i. The concavity

property implies that the more bandwidth a customer buys from the CSP, the cheaper the unit price is. Such pricing policy can incentivize cloud customers to buy more bandwidth. For a VSP, one of its objective is to minimize its total bandwidth cost, namely, $BC = \sum_{i=1}^{M} \psi_i(b_i)$. To take the user experience into account, we also associate the edges between data centers and regions with a cost, which represents the distribution latency from a data center to users in a region. Users prefer to receive video streams from data centers deployed in the same region to better their viewing quality. To this end, for each edge between nodes DC_i/V_j and R_k/V_j, the associated latency cost is assumed to be determined by a nondecreasing convex function $\omega_{ik}(\cdot)$. Intuitively, $\omega_{ik}(\cdot)$ stands for the latency between a data center DC_i and users in a region R_k, which is independent of distributed video V_j. In addition, $\omega_{ik}(\cdot)$ is a function of the amount of the bandwidth flow from a data center DC_i to users in region R_k. The intuition behind such definition is that the latency between a data center and a user region includes the propagation delay, the queuing delay, etc. The propagation delay is determined by the geographical distance, while the queuing delay is largely determined by the amount of the bandwidth flow according to the queuing theory. In spite that the propagation delay is independent of the amount of the bandwidth flow, the queuing delay will increase as a nondecreasing convex function of the amount of the bandwidth flow. The convexity property of the latency cost function follows the assumption made in [23], considering the increasing the queuing delay when the amount of the flow increases. Other edges are associated with a cost 0.

In addition to bandwidth cost, the provisioning strategy should also minimize overall latency cost at the same time. By incorporating both bandwidth cost and latency cost into the network flow model, we can transform the deployment problem into a min-cost network flow problem, in which the operational cost and the user experience are both taken into account. Before delving into the joint optimization of both costs, we first consider the optimization of bandwidth cost and latency cost individually.

Note that, different from computation-centric applications, video distribution is a bandwidth-centric service, in which bandwidth cost accounts for the majority of deployment cost compared with storage and computation cost. Given the small fraction of storage and computation cost in the deployment of video distribution services on the cloud, we focus more on optimizing bandwidth provisioning in this paper. Anyway, our framework model can be easily extended to take other types of resource costs into account.

C. Bandwidth Cost-Only Optimization

In the bandwidth cost-only optimization problem, denoted as **P1**, the objective of a VSP is to minimize the total bandwidth cost (denoted by BC), while satisfying the total user demands in all regions. All the latency costs $\omega_{ik}(\cdot)$ in the network flow model are set to zero. It applies to the case when a VSP is sensitive to the operational cost but regardless of the user experience. Based on the network flow theory, the problem **P1** can be formulated as follows:

$$\textbf{P1:} \quad \text{minimize} \quad BC = \sum_{i=1}^{M} \psi_i\Big(\sum_{k=1}^{K}\sum_{j=1}^{N} b_{ik}^{j}\Big)$$
$$\text{s.t.} \quad \sum_{i=1}^{M} b_{ik}^{j} \geq d_k^j, \forall k, j$$
$$b_{ik}^{j} \geq 0, \forall i, j, k. \tag{1}$$

By applying the nonlinear optimization theory [25], we can obtain Theorem 1.

Theorem 1: For the bandwidth cost-only optimization problem **P1**, the optimal bandwidth provisioning strategy is to only let the i^*th data center distribute all the videos, where $i^* = \arg\min_i\{\psi_i(D), i = 1, \cdots, M\}$, and $D = \sum_{k=1}^{K} d_k$.

Proof: See our technical report [26] for the proof details. ■

The intuition behind the above theorem is that, if not taking latency cost into account, a VSP prefers to only provision bandwidth at the data center with the lowest bandwidth cost and use one data center to serve video requests from all regions. For the implementation, as all bandwidth cost functions $\psi_i(\cdot)$ can be known beforehand, the problem **P1** can be easily solved by a central coordinator and we can find the data center with the lowest cost. Later, when receiving a video request, the coordinator returns the address of that data center to the user. However, the above solution is absolutely not a good strategy as the user experience directly impacts the market share of a video service provider.

D. Latency Cost-Only Optimization

For the latency cost-only optimization, a VSP aims at optimizing the user experience at all regions without considering the bandwidth cost. In this case, we set all the bandwidth cost $\psi_i(\cdot)$ associated with edges in the network flow model to zero. Then, the latency cost-only optimization problem **P2** can be formulated as follows:

$$\textbf{P2:} \quad \text{minimize} \quad LC = \sum_{i=1}^{M}\sum_{k=1}^{K} \omega_{ik}\Big(\sum_{j=1}^{N} b_{ik}^{j}\Big)$$
$$\text{s.t.} \quad \sum_{i=1}^{M} b_{ik}^{j} \geq d_k^j, \forall j, k$$
$$b_{ik}^{j} \geq 0, \forall i, j, k. \tag{2}$$

Accordingly, by solving the problem **P2**, we can obtain the following theorem.

Theorem 2: For the latency cost-only optimization problem **P2**, the optimal bandwidth provisioning strategy $\mathbf{B} = \{b_{ik}^{j}, \forall i, k, j\}$ is given by

$$b_{ik}^{j} = \begin{cases} d_k^j, & i = i^*(k) \\ 0, & i \neq i^*(k) \end{cases}$$

where $i^*(k) = \arg\min_i\{\omega_{ik}(d_k), i = 1, \cdots, M\}$.

Proof: See our technical report [26] for the proof details. ■

Intuitively, when only optimizing latency cost, the optimal provisioning strategy adopted by a VSP should allow each

viewer to receive its full video stream from the data center with the lowest latency. In an extreme case, suppose that there is one data center deployed at each region and the latency to a local data center is smaller than that to any remote data center, the optimal strategy to minimize overall user latency is to localize all the video traffic and only let local data centers serve users in the same region.

The provisioning strategies derived from **P1** and **P2** can be considered two extreme cases. In reality, the VSP needs to take both bandwidth cost and latency cost into account. However, these two objectives will be conflicting with each other. For example, when we try to reduce latency cost by provisioning more bandwidth at the local data center with a higher bandwidth price, the total bandwidth cost of a VSP will increase. Therefore, for a multiobjective optimization problem, we aim to find a set of Pareto points, in which no single objective can be improved without decreasing another. Among all the Pareto points, the optimal Pareto point guarantees both Pareto optimality and fairness.

IV. Joint Optimization for Cloud-Assisted Video Distribution

In this section, we consider to optimize the bandwidth cost and the latency cost jointly. To expose the tradeoff between two conflicting objectives, we first define the objective function as the simple weighted sum of the bandwidth cost BC and the latency cost LC, where α is a nonnegative scalar value that indicates the relative weight of two objectives

$$\begin{aligned} \textbf{P3:} \quad \text{minimize} \quad & BC + \alpha \cdot LC \\ \text{s.t.} \quad & \sum_{i=1}^{M} b_{ik}^{j} \geq d_k^j, \forall j, k \\ & b_{ik}^{j} \geq 0, \forall i, j, k. \end{aligned} \tag{3}$$

From the constraints of optimization problem **P1**, **P2**, and **P3**, we can find that their feasible solutions are the same. Without loss of generality, we define the feasible solution set as $\mathbb{B}$ and optimization variables $\mathbf{B} = \{b_{ik}^j, \forall i, k, j\} \in \mathbb{B}$. Actually, α can be removed from the objective function of **P3** if we multiply the unit latency cost by a factor of α in the problem formulation. Then, the objective function of Problem **P3** can be further simplified as $BC + LC$. In the following sections, we will use $BC+LC$ to replace the original objective function of **P3** for simplicity.

A. Nash Bargaining Solution for Joint Optimization Problem

To solve the above joint optimization problem with conflicting objectives, we adopt the Nash bargaining solution [27] as a theory tool, which can ensure optimality and fairness simultaneously. Optimality refers to Pareto optimality, which implies that the solution achieves the maximum reduction of the total cost starting from an initial point not on the Pareto curve (i.e., the disagreement point in the Nash bargaining process). While fairness guarantees that the tradeoff between two conflicting objectives are balanced, which means that cost minimization should benefit both objectives simultaneously.

In the bargaining process, we can imagine that there are two independent components with conflicting objectives who are trying to cooperate with each other to achieve an optimal and fair operation point. We can find the optimal point of $BC+LC$ on the Pareto curve derived from the problem **P3** by solving the following optimization problem **P4** (see below). The optimal solution obtained in **P4** by the Nash bargaining process corresponds to the optimal solution in **P3** since any further reduction of the total cost is not possible. In that case, we can achieve the minimum total cost

$$\begin{aligned} \textbf{P4:} \quad \text{maximize} \quad & |BC - BC_0| \cdot |LC_0 - LC| \\ \text{s.t.} \quad & \mathbf{B} \in \mathbb{B} \end{aligned} \tag{4}$$

where (BC_0, LC_0) is the disagreement point that represents the starting point of the negotiation. The problem **P4** aims at obtaining the optimal Pareto point that lies on the Pareto curve generated by solving **P3**. Denote BC_0 as the minimum bandwidth cost, and LC_0 as the maximum latency cost under the current demand, respectively. From the conflicting property of the bandwidth cost and the latency cost, the minimum bandwidth cost also implies the maximum latency cost. Therefore, we can guarantee that the point (BC_0, LC_0) also lies in the feasible solution set and can be chosen as the disagreement point. Therefore, the optimization problem can be rewritten as follows:

$$\text{maximize} \quad (BC - BC_0) \cdot (LC_0 - LC). \tag{5}$$

The maximization of $(BC - BC_0)(LC_0 - LC)$ is a transformation of the original optimization problem. By utilizing the Nash bargaining solution to obtain the optimal solution of $(BC - BC_0)(LC_0 - LC)$, we are able to obtain the optimal solution of the original optimization problem **P3**. From the conflicting property of BC and LC, if we decrease the value of BC, then the value of LC will be increased. The Nash bargaining solution can achieve the optimal point (BC, LC) that maximizes the value of $(BC - BC_0)(LC_0 - LC)$.

Define the total bandwidth supply from the ith data center to the kth region as $\hat{b}_{ik} = \sum_{j=1}^{N} b_{ik}^j$ and $\hat{\boldsymbol{b}} = \{\hat{b}_{ik}, \forall k, i\}$. Denote the bandwidth cost of the ith data center as $BC_i = \psi_i(\sum_{k=1}^{K} \hat{b}_{ik})$ and the latency cost associated with the ith data center as $LC_i = \sum_{k=1}^{K} \omega_{ik}(\hat{b}_{ik})$. By applying the optimization decomposition theory [28], we can derive a complexity-efficient algorithm. The optimization problem (5) can be rewritten as the following optimization problem since the log function does not change the optimal solution of the original problem:

$$\begin{aligned} \text{maximize} \quad & log(BC - BC_0) + log(LC_0 - LC) \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{M} \hat{b}_{ik} \geq d_k \\ & \hat{b_{ik}} \geq 0, \forall i, k. \end{aligned} \tag{6}$$

Then, we apply dual decomposition to the problem (6), and

get the following Lagrangian:

$$\begin{aligned} L(\hat{\boldsymbol{b}}, \boldsymbol{\lambda}) &= log(\sum_{i=1}^{M} \psi_i(\sum_{k=1}^{K} \hat{b}_{ik}) - BC_0) \\ &+ log(LC_0 - \sum_{i=1}^{M}\sum_{k=1}^{K} \omega_{ik}(\hat{b}_{ik})) \\ &+ \sum_{k=1}^{K} \lambda_k \cdot (\sum_{i=1}^{M} \hat{b}_{ik} - d_k) \end{aligned} \quad (7)$$

where $\lambda_k \geq 0$ is the Lagrange multiplier associated with the linear bandwidth supply constraint.

Before stepping into further analysis of dual decomposition, we should analyze the property of the Lagrangian (7). From the concavity of the bandwidth cost function and the convexity of the latency cost function, we can know that $(BC - BC_0)$ and $(LC_0 - LC)$ are both concave functions. The composition with the log function does not change their concave property. Therefore, the Lagrangian is also a concave function. The maximal value of the Lagrangian for a given λ_k is

$$\hat{\boldsymbol{b}}^{\star}(\boldsymbol{\lambda}) = \arg\max_{\hat{\boldsymbol{b}} \geq \mathbf{0}} [L(\hat{\boldsymbol{b}}, \boldsymbol{\lambda})]. \quad (8)$$

The result of (8) is unique due to the concavity of the Lagrangian. The dual problem of (6) is defined as follows:

$$\begin{aligned} \text{minimize} \quad & g(\boldsymbol{\lambda}) = L(\hat{\boldsymbol{b}}^{\star}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) \\ \text{s.t.} \quad & \boldsymbol{\lambda} \succeq \mathbf{0}. \end{aligned} \quad (9)$$

The dual problem (9) can be solved by subgradient method with the following updates:

$$\lambda_k(t+1) = [\lambda_k(t) - \beta(\sum_{i=1}^{M} \hat{b}_{ik}^{\star} - d_k)]^{+} \quad (10)$$

where β is a sufficiently small positive step-size. Nash bargaining solution can solve the joint optimization problem **P4** efficiently. Through dual decomposition, we can obtain optimal results by utilizing subgradient methods. A centralized algorithm of optimal bandwidth provisioning is provided in Algorithm 1.

In the centralized algorithm, the main complexity lies in the computation of the optimal result of (8). The updates of λ_k can be disseminated in parallel. In the practical implementation, the components to optimize the bandwidth cost and the latency cost can achieve the best tradeoff by iteratively negotiating the value of λ_k.

If the number of videos is huge, it will be impractical to place all videos at every data center. Based on the optimal bandwidth provisioning algorithm, we can derive the corresponding video placement strategy to minimize storage cost. The video placement algorithm is provided in Algorithm 2, which is inherently a greedy algorithm. In the algorithm, a video with higher popularity is first placed at a data center with the largest amount of available bandwidth for provisioning so that the number of replicas can be minimized (i.e., the storage space can be minimized).

In the case that prices of unit storage space are homogeneous across data centers, we have Theorem 3.

Algorithm 1 NBS-based Optimal Bandwidth Provisioning Algorithm

Input:

M, N, K;
Demand d_k;
Bandwidth cost function ψ_i;
Latency cost function ω_{ik};

Output:

The optimal bandwidth provisioning $\hat{b}_{ik}^{\star}$.

1: Initialization step: set $\lambda_k \geq 0$.
2: Compute the initial disagreement point (BC_0, LC_0) based on input.
3: Compute the optimal value of (8), then obtain optimal solutions $\hat{\boldsymbol{b}}^{\star}(\boldsymbol{\lambda}(t))$
4: Updates λ_k according to (10).
5: Set $t \leftarrow t+1$ and go to step 1(until satisfying termination criterion).
6: **Return:** $\hat{b}_{ik}$

Theorem 3: The greedy video placement algorithm shown in Algorithm 2 can minimize the overall storage cost.

Proof: See our technical report [26] for the proof details. ■

B. Online Algorithm for Joint Optimization

In the above section, we only consider the optimization of the joint cost function in the offline case, in which user demand in the future time slots can be known *a priori* by the algorithm. In this section, we extend our algorithm to a continuous online case, in which the algorithm has no knowledge about the future but use prediction methods to predict user demand in the future time slots. For each time slot, a VSP can adjust its deployment strategy at different data centers according to the predicted demand. To predict the demand of videos, we utilize prediction techniques proposed in previous work [9], [29], [30]. The design of a new prediction method is beyond the scope of this paper. In the online case, our objective is to minimize the joint cost function in each time slot and the probability of under-provisioning. At the beginning of each time slot, we first predict the demand of videos in all regions and then utilize Algorithm 1 to obtain the optimal bandwidth provisioning and video placement strategies.

Assume that the total demand from each region can be predicted. Let $\bar{d}_k$ denote the random variable that represents the actual total demand from the kth region. The mean and variance of $\bar{d}_k$ are μ_k and σ_k^2, respectively. For convenience, let $\boldsymbol{D} = [\bar{d}_1, \cdots, \bar{d}_K]$, $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_K]$ and $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_K]$. Assume that all random variables $\bar{d}_k$ follow Gaussian distributions $\bar{d}_k \sim \mathbf{N}(\mu_k, \sigma_k)$ (Note that the method can be easily extended to other distributions).

We can utilize the ARIMA and GARCH models [31] to calculate μ_k and σ_k based on the observed history of user demands. For example, suppose the next time slot to be t and denote the observed history of user demand from region k as $\boldsymbol{h_k} = \{d_k(0), d_k(1), \cdots, d_k(t-1)\}$, where $d_k(\tau), 0 \leq \tau \leq t-1$ is the observed user demand from region k by t. Then, μ_k in the time slot t can be calculated based on $\boldsymbol{h_k}$ by using the ARIMA model, and σ_k in the time slot t can be calculated

Algorithm 2 Greedy Video Placement Algorithm

Input:
Optimal bandwidth provisioning results $\hat{b}_{ik}^{\star}$(i.e., output from Algorithm 1)
Output:
Video placement strategy $\mathbf{1}_i^j$
Initialize $\mathbf{1}_i^j = 0, \forall i, j$
Sort data centers by their total bandwidth supply $b_i = \sum_{k=1}^{K} \hat{b}_{ik}^{\star}$ in the descending order, and generate a list of data centers L_d.
Sort videos by their total demand $\hat{d}^j = \sum_{k=1}^{K} d_k^j$, and generate a video list L_v.
Choose the head of the list L_v to be stored in $L_d(1)$.
Set $\mathbf{1}_{L_d(1)}^{L_v(1)} = 1$, $j = L_v(1)$, and $i = L_d(1)$.
for each region k **do**
  if the left demand of video j from region k $d_k^j > \hat{b}_{ik}^{\star}$ **then**
    Update the left available supply $\hat{b}_{ik}^{\star} = \hat{b}_{ik}^{\star} - d_k^j$, and the left demand $d_k^j = 0$.
  else
    Update the left demand $d_k^j = d_k^j - \hat{b}_{ik}^{\star}$, and the left available supply $\hat{b}_{ik}^{\star} = 0$.
  end if
end for
if the left demand of video $L_v(1)$ from each region is zero **then**
  Delete the head of L_v.
end if
Sort the list L_v by updated total demand in decreasing order.
if all the bandwidth supply of $L_d(1)$ have been used **then**
  Delete the head of L_d.
else
  Sort the list L_d by supply in the descending order.
end if
if L_d or L_v is empty **then**
  Return $\mathbf{1}_i^j$
else
  Back to step 2.
end if

based on $\boldsymbol{h_k}$ by using the GARCH model. Similar techniques are also used in [29] for the calculation of μ_k and σ_k.

To mitigate the probability of under-provisioning, the predicted demand input d_k should satisfy that the actual demand must be met with high probability

$$P(d_k < \bar{d}_k) \leq \epsilon$$

where ϵ is a small constant. Because $\bar{d}_k$ follows a Gaussian distribution, we have

$$d_k \geq \boldsymbol{E}[\bar{d}_k] + \theta\sqrt{\boldsymbol{var}[\bar{d}_k]} = \mu_k + \theta\sigma_k \tag{11}$$

where $\theta = F^{-1}(1-\epsilon)$.

In the online case, the predicted demand is represented by $d_k = \mu_k + \theta\sigma_k$. At the beginning of each time slot, we calculate (μ_k, σ_k) first so as to obtain the predicted demand d_k. The online bandwidth provisioning strategy is illustrated in Algorithm 3. Note that the prediction accuracy affects the performance of online algorithms. However, existing prediction techniques can guarantee that the prediction accuracy is very high for normal demand patterns (e.g., diurnal pattern of user demands). In our later experiments, it is found that the impact is not very significant even for a lower prediction accuracy.

Algorithm 3 Online NBS-Based Bandwidth Provisioning Algorithm

Input:
M, N, K;
μ_k and σ_k for random variable $\bar{d}_k$;
Bandwidth cost function ψ_i;
Latency cost function ω_{ik};
Output:
The optimal bandwidth provisioning $\hat{b}_{ik}^{\star}$.
Initialization step: set $\lambda_k \geq 0$.
Compute the predicted demand $d_k = \mu_k + \theta\sigma_k$.
Compute the initial disagreement point (BC_0, LC_0) based on predicted demand d_k.
Compute the optimal value of (7), then obtain optimal solutions $\hat{\boldsymbol{b}}^{\star}(\boldsymbol{\lambda}(t))$
Updates λ_k according to (10).
Set $t \leftarrow t+1$ and go to step 3(until satisfying termination criterion).
Return: $\hat{b}_{ik}$

C. Optimization for Video Distribution With Peer Contribution

Even in the context of cloud-assisted video distribution, the P2P technique is still useful to reduce the provisioning cost and increase system scalability. If taking resources contributed by end users into account, the deployment strategy of a VSP should be adjusted accordingly. To reduce distribution latency, we consider a locality-aware P2P distribution scenario, in which P2P traffic is confined to the same region where a viewer resides.

Although peer contribution can reduce the amount of provisioned bandwidth, the efficiency of peer exchange is impacted by the number of peers in the same distribution swarm. An efficiency function $\eta(x)$ is defined to represent the fraction of demand that can be satisfied by peer contribution, where x is the number of peers in the swarm. $\eta(x)$ is assumed to be a nondecreasing function, which implies that the efficiency of peer exchange can be improved with more peers in the swarm. $\eta(0) = 0$, $\eta(x) \in [0, 1]$ when $x > 0$. Suppose that there are n_k users in the kth region, if considering peer contribution, the optimization problem **P4** is transformed to the following problem **P5**:

$$\begin{aligned} \textbf{P5:} \quad \text{maximize} \quad & log(BC - BC_0) + log(LC_0 - LC) \\ \text{s.t.} \quad & \sum_{i=1}^{M} \hat{b}_{ik} \geq d_k(1 - \eta(n_k)) \\ & \hat{b_{ik}} \geq 0, \forall i, k. \end{aligned} \tag{12}$$

As to the number of future users in each region, it can be derived by the predicted demand directly. The Lagrangian (7)

is redefined as follows:

$$\begin{aligned} L(\hat{\boldsymbol{b}}, \boldsymbol{\lambda}) \quad = \quad & log(\sum_{i=1}^{M} \psi_i(\sum_{k=1}^{K} \hat{b}_{ik}) - BC_0) \\ & + log(LC_0 - \sum_{i=1}^{M}\sum_{k=1}^{K} \omega_{ik}(\hat{b}_{ik})) \\ & + \sum_{k=1}^{K} \lambda_k \cdot (\sum_{i=1}^{M} \hat{b}_{ik} - d_k(1 - \eta(n_k))). \end{aligned} \tag{13}$$

The subgradient update (10) is transformed into

$$\lambda_k(t+1) = [\lambda_k(t) - \beta(\sum_{i=1}^{M} \hat{b}^{\star}_{ik} - d_k(1 - \eta(n_k)))]^{+}. \tag{14}$$

The online NBS-based bandwidth provisioning strategy (Algorithm 3) can still be utilized to solve the problem **P5**. In the practical implementation, a tracker server can be deployed at each region to maintain peer information (similar approaches have been adopted by [7], [10]). A user can obtain video services from two resources: 1) cloud data centers, or 2) other peers in the same region. To better utilize the resources contributed by peers, a user will first query the tracker to obtain a peer list and seek to obtain video data from other peers in the same region; only when the request cannot be served by local peers will it be routed to a cloud data center.

V. Experimental Evaluation

In this section, we conduct a series of simulation experiments to evaluate the effectiveness of our proposed algorithms. To be more realistic, we first use the datasets obtain from PPLive to drive our simulation and compare with other alternative strategies. More details about simulation experiments is available in our technical report [26].

A. Dataset Description

The PPLive VoD dataset consists of over 25 million server-side log records on November 7, 2010. Each log record contains information, including the viewer ID, viewing starting time, viewing duration time, streaming rate, and the viewed video's ID. We divide the whole day into 48 time slots, each of which lasts for 30 min.

To obtain geographical distribution of viewers, we exploit MaxMind GeoIP [32] database to map users' IP addresses to six geographic regions, denoted by $R1, R2, \cdots, R6$. Note that the region $R6$ is defined as a virtual region that includes viewers whose IP addresses cannot be mapped successfully or from regions with too few viewers.

We set the locations of data centers based on the statistics reported by [33]. The bandwidth prices at each location are retrieved from the websites of major ISPs or cloud providers (e.g., China Telecom [34], Amazon EC2 [35], etc.) in the corresponding region. It is found that their pricing strategies usually bear concave property. Table I shows the prices to rent 100-Mbps bandwidth per day in each region. Due to the space limit, we have not listed the entire pricing plan of each region in our paper.

TABLE I
Prices to Rent 100-Mbps Bandwidth Per Day in Each Region (in the Unit of Dollars)

Region	R1	R2	R3	R4	R5	R6
Price	35.85	57.38	66.92	40.01	43.23	53.85

We assume that propagation latency is proportional to the geographical distance between two regions. Similar to [21], we define propagation latency as $0.02 * dist(km) + 5$, where *dist* is the geographical distance between any two regions and can be obtained from Google Map [36]. The geographical distance between the region $R6$ and other regions are assumed to be 1500 km. As to the efficiency of peer exchange, the efficiency function $\eta(x)$ is defined as a nondecreasing concave function where x is the number of viewers in a given region. In our experiments, $\eta(x)$ is defined as $\min\{\lfloor \frac{x}{300} \rfloor \cdot 0.1, 1\}$. The value of $\eta(x)$ is within the range [0, 1] in any region.

For comparison, we also consider three other provisioning strategies, which are listed as follows:

1) *centralized strategy*, in which all video requests are served by a single data center, which has the lowest bandwidth cost;
2) *local-only strategy*, in which video requests are only served by local data centers with the lowest latency;
3) *random strategy*, in which each video request is served by a random data center regardless of its location.

B. PPLive Trace-Driven Simulation

We first compare the normalized total cost $BC + LC$ of our proposed algorithms with other provisioning strategies in Fig. 3(a). Due to the demand dynamics, the curves fluctuate with time. It can be observed that NBS strategy can reduce around 80% of the total cost compared with the centralized and random strategy. The reduction is about 50% compared with the local-only strategy. By considering peer contribution, we can further reduce the incurred total cost.

To analyze our proposed algorithms in detail, we study the incurred the bandwidth cost and the latency cost separately. In order to facilitate the comparison among different strategies, we introduce a new metric called normalized cost ratio, which is defined as the cost incurred by the current strategy normalized by the sum of costs incurred by all strategies (including the current one). Fig. 3(b) illustrates the normalized bandwidth cost ratio under different bandwidth provisioning strategies. As expected, the local-only strategy incurs the largest bandwidth cost, and the optimal bandwidth cost can be achieved by the centralized strategy. The NBS strategy can reduce around 10% of the bandwidth cost compared with the local-only strategy, and the NBS-P2P strategy can further reduce 5%–15% bandwidth cost.

Fig. 3(c) illustrates the normalized latency cost ratio under different bandwidth provisioning strategies. A local-only strategy achieves the minimal latency cost. Our proposed NBS-P2P strategy can achieve comparable latency cost as that of the

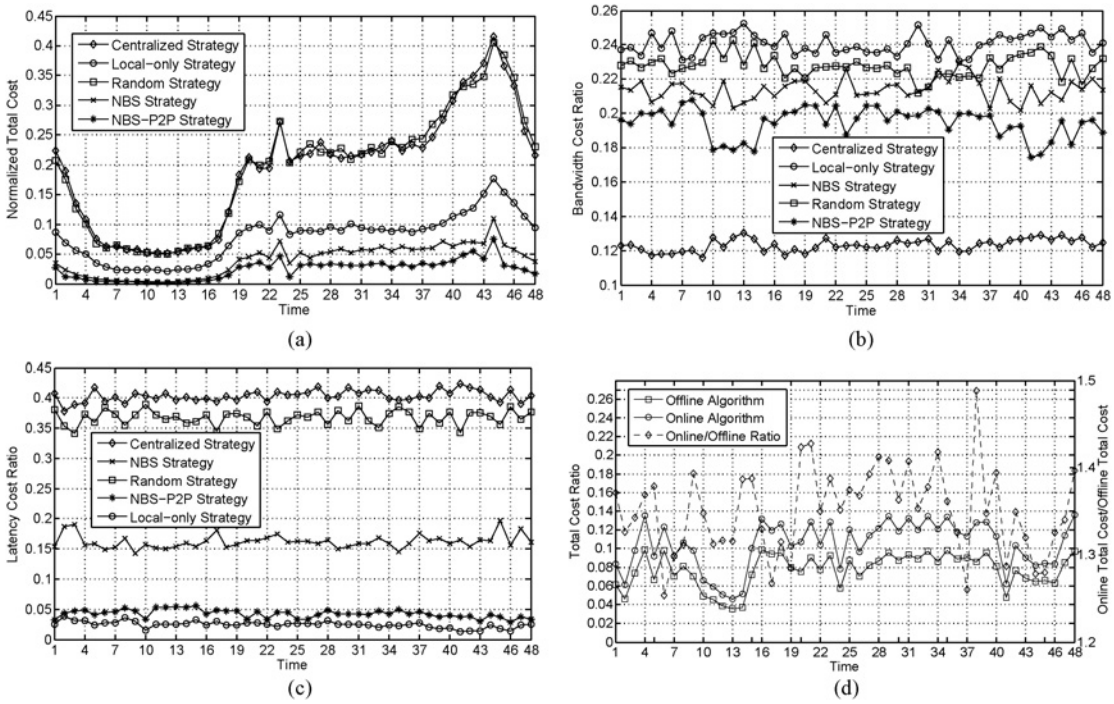


Fig. 3. Comparison of different bandwidth provisioning strategies (using PPLive trace-driven simulation). (a) Normalized total cost. (b) Normalized bandwidth cost ratio. (c) Normalized latency cost ratio. (d) Competitiveness of online algorithm.

local-only strategy. NBS-based strategies can reduce latency cost significantly, compared with centralized strategy and random strategy. More specially, the NBS strategy can reduce around 60% of latency cost compared with the centralized strategy, and the NBS-P2P strategy can reduce around 65% of latency cost compared with the NBS strategy.

We also evaluate the efficiency of our proposed online algorithm by comparing it with the offline algorithm. Fig. 3(d) shows that the efficiency ratio is within the range (1.25, 1.5).

C. Synthetic Trace-Driven Experiments

The main purpose to conduct synthetic trace-driven experiments is to validate the applicability of our algorithms in different scenarios, as the PPLive trace only represents one kind of user demand. In the section, we adopt more general cost functions and run a set of synthetic trace-driven experiments in a larger scale.

In our experiments, we consider a synthetic scenario in which a video service provider relies on 30 geo-distributed data centers on the cloud platform to distribute videos to users spreading over 50 regions. The bandwidth cost function associated with the ith data center is defined as $\psi_i(x) = \rho_i \cdot x^{\gamma}$, where ρ_i is a scalar and $\gamma \in (0, 1)$ is a factor to guarantee the concavity of $\psi_i(x)$. The latency cost function between a data center and a region is defined as $\omega(x) = \mu \cdot x^{\nu}$, where μ is a scalar to identify different cost functions for every data center and region pair and ν is used to ensure the convex property of $\omega(x)$. For a local data center and region pair, the factor μ will be less than that of a remote pair. To simplify our experiments, ρ_i is defined as a random variable that is uniformly distributed in the range of [1,20], denoted by $\rho_i \sim U(1, 20)$.[1] We also assume that $\gamma = 0.5$, $\mu \sim U(1, 5)$, $\nu = 1.5$ in the experiments. However, it should be noted that any concave bandwidth cost function and convex latency cost function are applicable in the experiments. As to the efficiency of peer exchange, we define the function $\eta(x)$ in a similar way as that in the trace-driven experiments.

We follow a similar approach as [9] to generate synthetic video demands in each region. All regions are divided into three classes, which represent user population in different scales. The number of regions in each class and their corresponding parameters are given as follows:

1) *Class I*: 20 regions, with mean $\mu_i \sim U(100, 500)$ and variance $\sigma_i \sim U(10, 50)$;
2) *Class II*: 20 regions, with mean $\mu_i \sim U(500, 1000)$ and variance $\sigma_i \sim U(10, 100)$;
3) *Class III*: 10 regions, with mean $\mu_i \sim U(1000, 1500)$ and variance $\sigma_i \sim U(10, 150)$.

Similar to the above section, we evaluate the efficacy of different bandwidth provisioning algorithms by comparing the total cost $BC + LC$. Fig. 4(a) illustrates the normalized total cost resulted from different bandwidth provisioning strategies. If not taking peer contribution into account, the NBS strategy achieves the optimality in all time slots. Due to the convex property of latency cost function, local-only strategy can obtain sub-optimal results compared with centralized strategy and random strategy. The NBS strategy can reduce about 49% of the total cost compared with local-only strategy. We also observe that the NBS-P2P strategy can reduce about 51% of the total cost compared with the NBS strategy due to bandwidth resources contributed by peers.

[1] $U(x, y)$ means a uniform distribution in the range $[x, y]$.

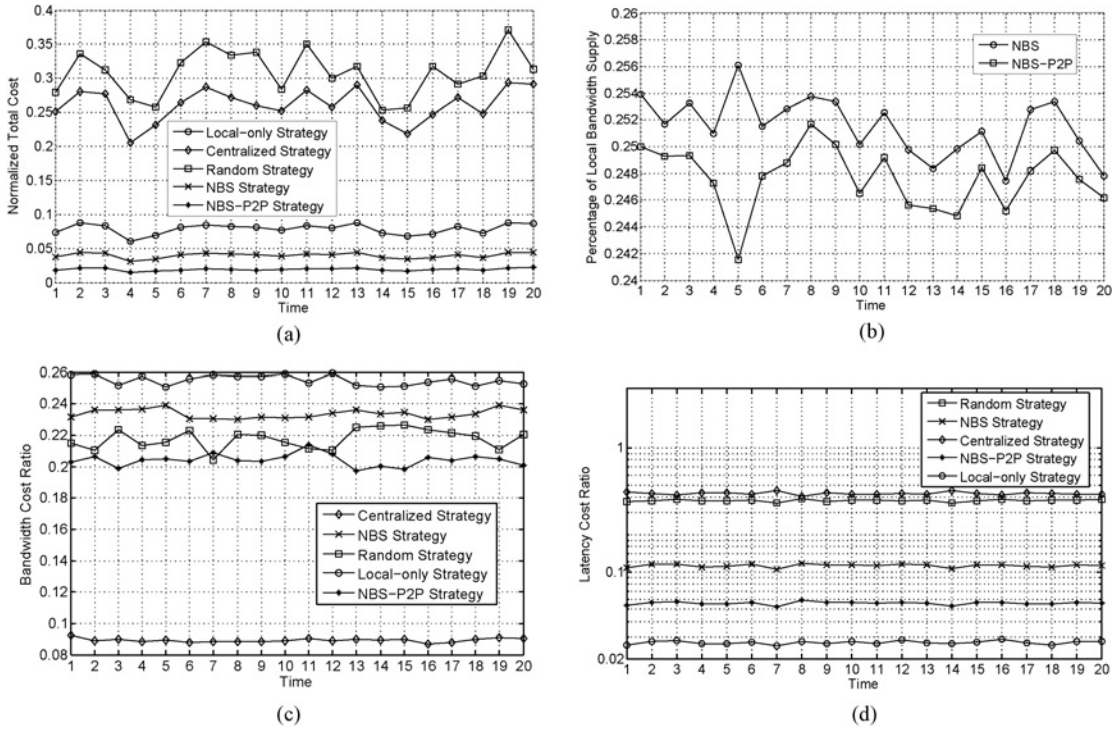


Fig. 4. Comparison of different bandwidth provisioning strategies (using synthetic trace-driven simulation). (a) Normalized total cost. (b) Percentage of local bandwidth supply in NBS-based strategies. (c) Normalized bandwidth cost ratio. (d) Normalized latency cost ratio.

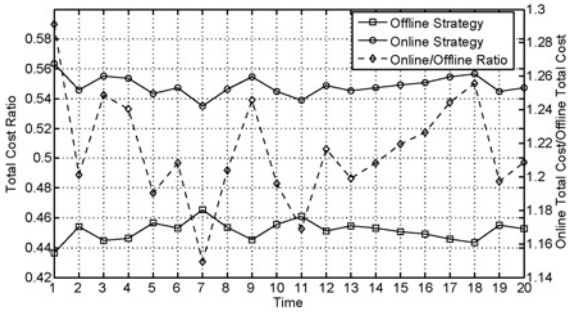


Fig. 5. Efficiency of online algorithm with prediction.

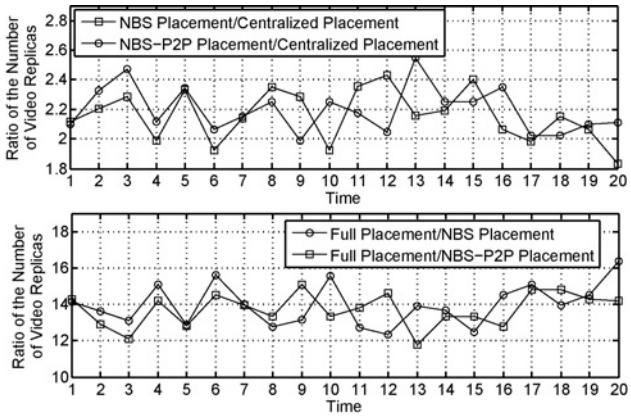


Fig. 6. Efficiency of video placement strategy.

We also analyze the percentage of local and remote bandwidth supply to understand the optimality of our proposed NBS-based algorithm. Fig. 4(b) shows the percentage of local bandwidth supply by all data centers in each time slot corresponding to two NBS-based algorithms. Here, NBS refers to Algorithm 3, while NBS-P2P refers to Algorithm 3 considering peer contribution. The ratio of local bandwidth supply of the NBS strategy is within the range (0.248, 0.256). By taking peer contribution into account, the ratio of local bandwidth supply can be further reduced to the range (0.242, 0.252).

To analyze the bandwidth and latency cost resulted from different bandwidth provisioning strategies, we calculate the bandwidth cost and the delay cost separately. Fig. 4(c) illustrates the normalized bandwidth cost ratio of different provisioning strategies. The centralized strategy incurs the least bandwidth cost, which well confirms the correctness of Theorem 1. On the contrary, the local-only strategy performs the worst in terms of bandwidth cost reduction. We observe that the NBS strategy can reduce around 10% of the bandwidth cost compared with the local-only strategy. Furthermore, the NBS-P2P strategy can reduce 5%–20% of the bandwidth cost compared with the NBS strategy.

Fig. 4(d) shows the normalized latency cost ratio of different provisioning strategies. Local-only strategy performs the best in terms of latency cost reduction, which confirms the correctness of Theorem 2. For latency cost, the reduction of two NBS-based strategies is close to the best strategy. Compared with the worst strategy, i.e., centralized strategy, the NBS strategy can reduce more than 70% of latency cost. Furthermore, the NBS-P2P strategy can reduce around 50% of latency cost compared with the NBS strategy.

Fig. 5 illustrates the competitiveness of our online NBS-based bandwidth provisioning algorithm compared with the

offline algorithm. In the offline algorithm, the demands can be known beforehand at the beginning of each time slot; while in the online algorithm, we need to predict the demand and use the predicted demand as the input. We also present the ratio between the costs incurred by online strategy and offline strategy in Fig. 5. In most cases, the cost of online algorithm is only 1.15–1.3 times that of offline algorithm.

To check the efficiency of our video placement algorithm, we also compare our algorithm (Algorithm 2) with two other video placement strategies: 1) *full placement* that places all videos in each data center, and 2) centralized placement that only places all videos in a single data center. We calculate the number of video replicas resulted from different strategies accordingly. Fig. 6 illustrates the ratio between the number of replicas resulted from different algorithms. For the NBS strategy, its competitiveness is 1.8–2.4 compared with centralized placement strategy, while the full placement strategy incurs around 15 times the number of replicas resulting from NBS strategy. In our experiment, we find that the number of videos stored in a data center will not exceed 20% of the total number of videos.

VI. Conclusion

The on-demand self-service feature of cloud computing enabled a video service provider to provision its cloud resources adaptively for video distribution. In this paper, we studied the optimal deployment of cloud resources in multiple geographically distributed data centers to improve the user experience and minimize the operational cost simultaneously. We built a network flow model to formulate the cloud deployment problem and derive corresponding optimal bandwidth provisioning and video placement strategies with Nash bargaining solutions. We also examined the case when peer contribution is considered. Our work in this paper provided useful guidelines for different video service providers to provision their services effectively. In the future, we plan to investigate the deployment over a hybrid cloud infrastructure with both public clouds and private clouds, and study how to better select data centers to optimize video applications with different QoS requirements.

Acknowledgment

The authors would like to thank the PPTV Company for kindly sharing the PPLive traces.

References

[1] Cisco Systems, Inc. (2011). *Cisco Visual Networking Index: Forecast and Methodology, 2010–2015* [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html

[2] M. Freedman, "Experiences with CoralCDN: A five-year operational view," in *Proc. USENIX NSDI*, 2010, pp. 7–7.

[3] P. Wendell and M. Freedman, "Going viral: Flash crowds in an open CDN," in *Proc. ACM IMC*, 2011, pp. 549–558.

[4] M. Cha, H. Kwak, P. Rodriguez, Y. Yeol Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the worlds largest user generated content video system," in *Proc. 5th ACM/USENIX Internet Meas. Conf.*, 2007, pp. 1–14.

[5] 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. (2009, Feb.). "Above the clouds: A Berkeley view of cloud computing," EECS Dept., Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2009-28 [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[6] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing: Directions and applications," *Signal Process. Mag.*, vol. 28, no. 3, pp. 59–69, 2011.

[7] Y. Wu, C. Wu, B. Li, X. Qiu, and F. Lau, "CloudMedia: When cloud on demand meets video on demand," in *Proc. IEEE ICDCS*, Jun. 2011, pp. 268–277.

[8] H. Li, L. Zhong, J. Liu, B. Li, and K. Xu, "Cost-effective partial migration of VoD services to content clouds," in *Proc. IEEE Int. Conf. Cloud Comput.*, Jul. 2011, pp. 203–210.

[9] D. Niu, H. Xu, B. Li, and S. Zhao, "Quality-assured cloud bandwidth auto-scaling for video-on-demand applications," in *Proc. IEEE INFOCOM*, vol. 12. Mar. 2012, pp. 460–468.

[10] F. Wang, J. Liu, and M. Chen, "CALMS: Cloud-assisted live media streaming for globalized demands with time/region diversities," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 199–207.

[11] Z. Huang, C. Mei, L. Li, and T. Woo, "CloudStream: Delivering high-quality streaming videos through a cloud-based SVC proxy," in *Proc. IEEE INFOCOM Mini Conf.*, Apr. 2011, pp. 201–205.

[12] F. Chen, K. Guo, J. Lin, and T. Porta, "Intra-cloud lightning: Building CDNs in the cloud," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 433–441.

[13] Y. Jin, Y. Wen, G. Shi, G. Wang, and A. Vasilakos, "CoDaaS: An experimental cloud-centric content delivery platform for user-generated contents," in *Proc. IEEE Int. Conf. Comput. Networking Commun.*, Jan. 2012, pp. 934–938.

[14] F. Liu, B. Li, L. Zhong, B. Li, H. Jin, and X. Liao, "Flash crowd in P2P live streaming systems: Fundamental characteristics and design implications," *IEEE Trans. Parallel Distributed Syst.*, vol. 23, pp. 1227–1239, 2012.

[15] Xunlei Inc. (2013). *Xunlei Kankan* [Online]. Available: http://www.xunlei.com/

[16] H. Yin, X. Liu, T. Zhan, V. Sekar, F. Qiu, C. Lin, H. Zhang, and B. Li, "LiveSky: Enhancing CDN with P2P," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, no. 3, p. 16, 2010.

[17] V. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z. Zhang, "Unreeling Netflix: Understanding and improving multi-CDN movie delivery," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1620–1628.

[18] V. Aggarwal, X. Chen, V. Gopalakrishnan, R. Jana, K. Ramakrishnan, and V. Vaishampayan, "Exploiting virtualization for delivering cloud-based IPTV services," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2011, pp. 637–641.

[19] D. Niu, C. Feng, and B. Li, "A theory of cloud bandwidth pricing for video-on-demand providers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 711–719.

[20] J. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, "The little engine (s) that could: Scaling online social networks," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 375–386, 2010.

[21] Y. Wu, C. Wu, B. Li, L. Zhang, Z. Li, and F. Lau, "Scaling social media applications into geo-distributed clouds," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 684–692.

[22] X. Cheng and J. Liu, "Load-balanced migration of social media to content clouds," in *Proc. ACM NOSSDAV*, 2011, pp. 51–56.

[23] W. Jiang, R. Zhang-Shen, J. Rexford, and M. Chiang, "Cooperative content distribution and traffic engineering in an ISP network," in *Proc. ACM SIGMETRICS*, Jun. 2009, pp. 239–250.

[24] D. Goldenberg, L. Qiu, H. Xie, Y. Yang, and Y. Zhang, "Optimizing cost and performance for multihoming," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 4, pp. 79–92, 2004.

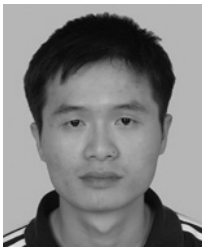
[25] D. Bertsekas, *Convex Optimization Theory*. Athena Scientific, Nashua, NH, USA, 2009.

[26] J. He, D. Wu, Y. Zeng, X. Hei, and Y. Wen. (2012). "Toward optimal deployment of cloud-assisted video distribution services," CS-2012-0616, Sun Yat-Sen University, Tech. Rep. [Online]. Available: http://sist.sysu.edu.cn/dwu/cloud-tr.pdf

[27] K. Binmore, A. Rubinstein, and A. Wolinsky, "The nash bargaining solution in economic modelling," *RAND J. Econ.*, vol. 17, no. 2, pp. 176–188, 1986.

[28] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Select. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[29] D. Niu, Z. Liu, B. Li, and S. Zhao, "Demand forecast and performance prediction in peer-assisted on-demand streaming systems," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 421–425.
[30] G. Gursun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 16–20.
[31] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Berlin, Germany: Springer, 2002.
[32] MaxMind Inc. (2013). *Maxmind GeoIP* [Online]. Available: http://www.maxmind.com/app/country
[33] G. Zhang, X. Zhao, X. Hei, and W. Cheng, "A comparison measurement study of hybrid CDN-P2P VoD systems," EIE Dept., Huazhong Univ. Sci. Technol., Tech. Rep., 2012.
[34] China Telecom Inc. (2013). *China Telecom Homepage* [Online]. Available: http://www.chinatelecom.com.cn/.
[35] Amazon Inc. (2013). *Amazon EC2 Pricing* [Online]. Available: http://aws.amazon.com/ec2/pricing/
[36] Google Inc. (2013). *Google Map Homepage* [Online]. Available: http://maps.google.com/



Jian He received the B.S. degree in computer science from Sun Yat-Sen University, Guangzhou, China, in 2011.

He is currently a Graduate Student at the School of Information Science and Technology, Sun Yat-Sen University. His current research interests include content distribution networks, data center networking, green networking, and network measurement.

Di Wu (M'06) received the B.S. degree from the University of Science and Technology of China. Hefei, China, in 2000, the M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2003, and the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2007.

From 2007 to 2009, he was a Post-Doctoral Researcher with the Department of Computer Science and Engineering, Polytechnic Institute of NYU, Brooklyn, NY, USA, advised by Prof. K. W. Ross. He has been an Associate Professor with the Department of Computer Science, Sun Yat-Sen University, Guangzhou, China, since July 2009. His current research interests include multimedia communications, cloud computing, peer-to-peer networking, Internet measurement, and network security.

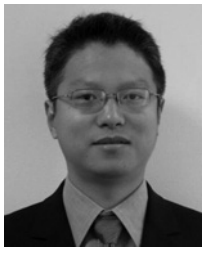
Mr. Wu was a recipient of the IEEE INFOCOM 2009 Best Paper Award, and is a member of the IEEE Computer Society, the ACM, and the Sigma Xi.

Yupeng Zeng received the B.Eng. degree in computer science and technology from Sun Yat-Sen University, Guangzhou, China.

Currently, he is a Graduate Student at Sun Yat-Sen University, working under the supervision of Dr. D. Wu. His current research interests include data center and cloud computing.

Xiaojun Hei (M'08) received the B.Eng. degree in information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1998, and the M.Phil. degree in electrical and electronic engineering from the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, in 2000, and the Ph.D. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology in 2008.

Since October 2008, he has been with the Internet Technology and Engineering Research and Development Center, Department of Electronics and Information Engineering, Huazhong University of Science and Technology. He is currently an Associate Professor with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology. From September 2005 to September 2007, he conducted a research visit on P2P networking with the Department of Computer Science and Engineering, Polytechnic Institute of NYU, Brooklyn, NY, USA. He is a co-author, together with Y. Liu and K. W. Ross, of the Best Paper in multimedia communications for 2008 awarded by the Multimedia Communications Technical Committee of the IEEE Communications Society. His current research interests include Internet measurement, applications, and architecture.



Yonggang Wen received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008.

He is currently an Assistant Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. Previously, he was with Cisco, San Jose, CA, USA, as a Senior Software Engineer and a System Architect for content networking products. He has also been a Research Intern with Bell Laboratories, Murray Hill, NJ, USA, Sycamore Networks, Chelmsford, MA, USA, and a Technical Advisor to the Chairman at Linear A Networks, Inc., Milpitas, CA, USA. His current research interests include cloud computing, mobile computing, multimedia networks, cyber security, and green ICT.