This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

Cost-Aware Traffic Management under Demand Uncertainty From a Colocation Data Center User's Perspective

Yong Zhan, Student Member, IEEE, Mahdi Ghamkhari, Member, IEEE, Hossein Akhavan-Hejazi, Member, IEEE, Du Xu, *Member, IEEE,* and Hamed Mohsenian-Rad, *Senior Member, IEEE*

Abstract—Burstable billing is widely adopted by colocation data center providers to charge their users for data transferring. This paper propose a cost-aware traffic management approach for a colocation data center user under burstable billing where it is charged based on the 95th percentile bandwidth usage. To do this, we first develop a tractable mathematical expression to calculate the *95th percentile usage* of a user. Then, we develop an optimization problem to maximize the user's surplus based on both deterministic and stochastic predictions of the user's demand. We show that the resulted optimization problem, while non-convex by nature, can be efficiently solved or approximated using a convex program. We also show that the proposed approach can also be applied in a more general scenario where the user gets services from multiple service providers. Using real-world workload traces, we show that the proposed approach can reduce a colocation data center user's IP transit cost by 26% and increase its total surplus by 23%, compared to the current practice of allocating bandwidth on-demand.

Index Terms—Burstable billing, bandwidth, demand uncertainty, nonlinear mixed-integer programming, surplus maximization.

1 Introduction

Bandwidth cost has become the second largest aspect of Data Centers (DCs) overall costs, second to energy cost, reported by Colocation America [1]. Cisco forecast that annual global DC IP traffic will reach 15.3 zettabytes by 2020, rising from 4.7 zettabytes per year in 2015 [1]. Nearly 23% of the overall DCs traffic, i.e., the data transferring from/to DCs, is usually charged by a smart data pricing method called burstable billing [1]. Burstable billing is used in practice, by Internet Service Providers and Colocation Data Center (CDC) providers to charge for transferring data. [2], [3], [4], [5], [6]. Burstable billing is also widely adopted by CDC providers such as The ANLX Server Farm[2], as a means to charge their DC users, for bandwidth usages.

In burstable billing mechanism a CDC provider, here after *provider*, who supports its *users*, i.e. the DCs, with links for data transferring, measures each user's bandwidth usage based on the user's peak usage at a certain percentile, often at the *95th percentile usage* [7]. By construction, burstable billing neglects the user's usage of bandwidth during any time other than periods of peak use. Hence, burstable billing allows users to exceed their usage thresholds for a short period without facing financial penalty [4]. An example for calculating the 95th percentile usage is shown in Fig. 1.

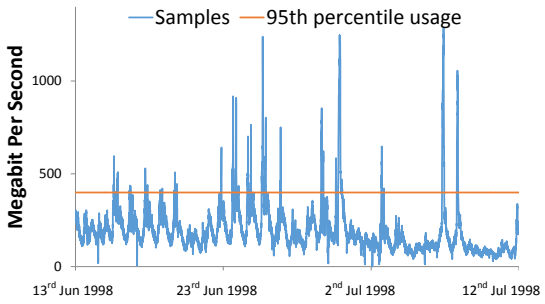


Fig. 1. An example for calculating the *95th percentile usage*: a total of 8640 samples are collected for a user during one billing cycle. After throwing away the top 5%, i.e., $5 * 8640/100 = 432$ samples, the *95th percentile usage* is obtained as 399.1277 Mbps, which is equal to the highest recorded bandwidth usage of the remaining $95 * 8640/100 = 8208$ samples. The *95th percentile usage* is shown by the red line. Here, the user is allowed to have a total of 432 bursts above the red line without facing financial penalty.

1.1 Motivation

In this paper, as illustrated by Fig. 2, we are interested in studying burstable billing from the CDC user's viewpoint . Large volumes of data transferring leads to extremely high IP transit cost for users. Accordingly, how to reduce the user IP transit cost has become a big concern [8], [9], [10].

A common strategy for a user to reduce its IP transit cost, is to move its traffic across time to avoid coinciding peak usage, thus, reducing the overall peak or/and 95th percentile usage of bandwidth. To get a sense of cost-aware traffic

- *Y. Zhan and D. Xu are with the Key Laboratory of Optical Fiber Sensing and Communications, University of Electronic Science and Technology of China, Chengdu, China.*
  E-mail: {yzhan.china, xudu.uestc}@gmail.com.
- *M. Ghamkhari, H. Akhavan-Hejazi and H. Mohsenian-Rad are with the Department of Electrical Engineering, University of California, Riverside, CA, USA.*
  E-mail: {ghamkhari, shejazi, hamed}@ece.ucr.edu.

This work was done when Y. Zhan was a Visiting Student at the University of California at Riverside. This work is supported in part by NSF grant 1319798. The corresponding author is H. Mohsenian-Rad.

1. http://www.colocationamerica.com/data-center-connectivity/bandwidth.htm.
2. http://www.anlx.com/server-hosting/rack-space/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

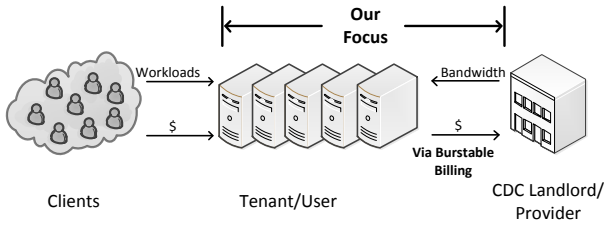


Fig. 2. An example setup for the application of burstable billing: a tenant in a colocation data center who serves outside clients with uncertain demands.

management, consider a user who is charged by burstable billing based on its 95th percentile usage. In this case, if the user decides to use bandwidth on-demand, it's 95th percentile usage equals 326.27Mbps (as shown in Fig. 3(a)). With a simple traffic management though, such as deferring workload as shown in Fig. 3(b), the user's 95th percentile usage can be reduced to 235.52Mbps . The detail of traffic adjustment is shown in Fig. 3. However, considering that traffic management may require extra O&M operations or/and lead to performance degradation, e.g., increase of latency, whether or not DCs are willing to modulate their traffic is often overlooked. For instance, 100ms of increase of latency can cost Amazon 1% loss in sale [11].

In this paper we address the trade-off between cost and performance based on user's preferences. Specifically, we seek to answer this fundamental question: *What is the best way for an individual CDC user who is charged via burstable billing, to manage its operation and the use of bandwidth*? Our approach to answer this question is based on formulating and solving an optimization problem for bandwidth usage which aims at maximizing the user's *surplus*, i.e., its net utility minus cost. A hindrance so far, in accurate tractable modeling, and optimization of the trade-off between utility and cost of the user traffic management has been lied on the tractable modeling of the 95th percentile usage cost within the optimization. We address this issue in the current work.

We take into consideration the fact that, in practice, neither the user nor the provider have perfect knowledge about the traffic, and thus the demand for bandwidth is unknown. For example, when it comes to a user in a CDC as in Fig. 2, it has no idea when and how many requests it will receive from its clients, and thus it cannot perfectly predict its traffic in the future. Therefore, in our analysis, we address demand uncertainty within a stochastic optimization framework.

1.2 Contributions

The main contributions of this paper are as follows:

1) We develop a tractable optimization-based model to obtain the user's strategy for traffic management under burstable billing. Our model includes the user's cost based on the 95th percentile usage of bandwidth, as well as the user's surplus model, user's cost and utility, in the bandwidth allocation problem.
2) The traffic management optimization problem proposed in this work, maximizes the user's surplus based on both deterministic, e.g., with a single profile, and stochastic, e.g., with various statistics, predictions of the user's demand. Our proposed model takes into account the uncertainty of bandwidth demand with arbitrary probability distributions at the time of decision.
3) We show that the optimization problems, while non-convex by nature, can be efficiently solved or approximated using several convex programming techniques. We extend the models and solutions also to a more general scenario where users get services from multiple service providers.
4) We evaluate our design based on a real-world workload trace: Wikipedia Page View data [12]. With a typical workload forecasting method, we show that the use of our design is particularly rewarding if a user is charged by high bandwidth price and/or it is more sensitive to price than to performance.

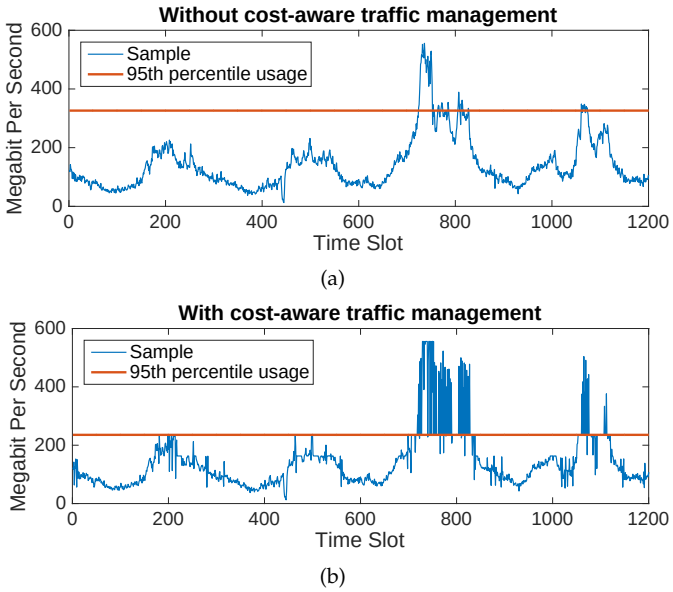


Fig. 3. Cost-aware traffic management, a use scenario; a) without cost-aware traffic management, b) with cost-aware traffic management,.

2 Related Work

Traffic management under burstable billing can be studied from two different viewpoints: providers and users. While there are more previous studies on analysis of burstable billing from the perspective of providers, there have been few studies from the viewpoint of users and how they optimally respond and manage their traffic in response to burstable billing. For studies that address burstable billing from the providers viewpoint, e.g. in [3], [4], [8], [13], a common strategy is for the provider to move different users' workloads across space and time to avoid coinciding their peak usages, thus, reducing the overall peak demand for bandwidth [13]. However, whether or not users are willing to modulate their workloads is often overlooked. In contrast, our focus in this paper, is to address the traffic management from a user's viewpoint and the optimal user response under burstable billing.

The studies that address burstable billing from the user's perspective have emerged only recently. Among those, there are works that perform various data driven or statistical

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

analysis on user traffic and/or costs under burstable billing. The authors in [3] analysed the impact of the length of time interval. They find out that a user's 95th percentile usage is usually inversely proportional to the length of time interval, i.e., with the increasing of the length of time interval, the user's 95th percentile usage convexly decreases.

For the studies that propose methods for user traffic management under burstable billing, so far due to the lack of an optimization-based mathematical model for the 95th percentile usage cost of the bandwidth, a common approach in studies such as [6], has been to use experimental and/or heuristic methods, where as in this work we rather focus on developing an analytical model for user traffic management. There are also few studies that are analytical; however, they still make assumptions to avoid modeling of the 95 percentile billing cost. For example, they develop models based on the 100 percentile billing (i.e., peak pricing) instead of 95 percentile billing [14], or they assume that the cost of bandwidth is volume-based [15], [16], or assume that the workload has a specified distribution, e.g., Gaussian distribution [17].

Additionally, previous studies on burstable billing from the user's viewpoint, e.g. [7], [9], [18], [19], [20], [21], have not accounted for performance degradation due to traffic adjustment, thus, both incentives and extra costs for the traffic adjustment are not considered simultaneous in the models. For example, studies in [14], [19], [20], [22] try to reduce a user's 95th percentile usage or peak usage of bandwidth via postponing parts of the user's traffic. These studies neglect that the users may not tend to adjust their traffic if they are not well incentivized since traffic management may cause extra operation cost and/or performance degradation, e.g., increase of latency. In this paper, we take into consideration the utility loss caused by traffic management and aim at maximizing a CDC user's surplus, i.e., minimizing its utility loss and IP transit cost.

3 Modeling User's Cost and Surplus under Burstable Billing

In this section, we obtain mathematical expressions to model a user's bandwidth cost, bandwidth revenue, and net surplus under burstable billing, i.e when users is charged based on *95th percentile usage*, of bandwidth.

3.1 Bandwidth Usage Cost

In burstable billing, a provider divides a billing cycle into τ time intervals of equal length T. We assume the typical interval length of of $T = 5$ minuets [3]. The provider takes samples of the user's bandwidth usage , e.g., once every five minutes during that billing cycle. Let $x[1], \cdots, x[\tau]$ denote the user's bandwidth usage samples in time slots $t = 1, \cdots, \tau$. The provider charges the user at certain rate based on the *95th percentile usage* of bandwidth. Let δ ($/Mbps) denotes the price of bandwidth. We assume that the price of bandwidth in a billing cycle is constant. We also denote the 95th percentile usage of bandwidth by $\mu_{95}(x[t])$. Accordingly, the bandwidth cost based on burstable billing, C_{95}, for the user given the bandwidth usage samples $x[1], \ldots, x[\tau]$ is expressed as

$$C_{95}(x[t]) = \delta \cdot \mu_{95}(x[t]), \tag{1}$$

To obtain the user's *95th percentile usage*, the top 5% of the samples gathered within the billing cycle are thrown away and the highest element of the remaining 95% samples is taken as the user's *95th percentile usage*. An example for calculating the $\mu_{95}(x[t])$ is shown in Fig. 1. The user's 95th percentile usage, $\mu_{95}(x[t])$, given the bandwidth usage samples $x[1], \cdots, x[\tau]$ can be mathematically expressed as

$$\begin{aligned} \mu_{95}(x[t]) = \min_{\rho} \max_{t} \rho[t]x[t] \\ \text{s.t. } \rho[t] \in \{0,1\}, \quad \forall t, \\ \sum_{t=1}^{\tau} \rho[t] = \lceil 0.95\tau \rceil, \end{aligned} \tag{2}$$

where $\lceil \cdot \rceil$ denotes the ceiling function. The variables in the above minimization are $\rho[t]$ for all $t = 1, \ldots, \tau$. For each sample $x[t]$, if $\rho[t] = 0$, it indicates that its corresponding usage $x[t]$ is within the top 5% of the values in $x[1], \ldots, x[\tau]$ and thus the usage $x[t]$ at this time slot has no impact on the *95th percentile usage* $\mu_{95}(x[t])$. In other words, if $\rho[t] = 0$, the user can always utilize bandwidth on-demand without worrying about its bandwidth cost. On the contrary, if $\rho[t] = 1$, the user may restrict its usage at this time slot to reduce its *95th percentile usage*.

3.2 Bandwidth Usage Surplus

Next, we obtain the user's net surplus *prior* to a billing cycle. Let $D[t]$ (Mbps) be the user's demand for bandwidth at time slot t, which is the amount of bandwidth needed by the user to fully satisfy its clients. Note that, the user may not know its exact demand in the future, rather has a distribution for its demand, i.e., $D[t]$ is a random variable. We also assume that the user gains a *utility*, e.g. a revenue, from the bandwidth usage. Here, as in [23], [24], we assume a general net utility function that depends only on user's bandwidth usage. The utility function $U(\cdot)$ is a concave and non-decreasing function of the total bandwidth.

At each time slot, the user obtains the highest utility, when it fully serves the bandwidth demand, i.e., $D[t]$. However, the user may not always choose to serve the demand in full. Let $X[t]$ (Mbps) be the pre-cycle *planned* usage of bandwidth for the user during time interval $t = 1, \ldots, \tau$. Here, $X[t]$ is decided based on the demand $D[t]$. Accordingly, we can formulate the user's revenue in a billing cycle, that is corresponding to planned usage samples $X[1], \cdots, X[\tau]$ as

$$\sum_{t=1}^{\tau} U(\min\{X[t], D[t]\}). \tag{3}$$

From (1) and (3), the user's surplus is obtained as

$$S = \sum_{t=1}^{\tau} U(\min\{X[t], D[t]\}) - \delta \cdot \mu_{95}(X[t]). \tag{4}$$

4 Surplus Optimization Problem in Response to Burstable Billing

Typically, neither the user nor the provider have perfect knowledge about the user's bandwidth demand

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

in an upcoming billing cycle. Accordingly, parameters $D[1], \ldots, D[\tau]$ are often uncertain. Since the scope of this paper does not include workload forecasting, we assume that the prediction of user's demand $D[t]$ is given. Such prediction is either deterministic or stochastic. In this section, we formulate the optimization problems to maximize the user's surplus *prior* to a billing cycle under deterministic and stochastic prediction of $D[t]$.

4.1 Optimization Problem with Deterministic Prediction

If the prediction of demand for bandwidth is deterministic, i.e., parameters $D[1], \ldots, D[\tau]$ are *deterministic*, from (2) and (4), we formulate the optimization problem to maximize the user's surplus over a billing cycle as:

$$\begin{aligned} \max_{X[t],\rho[t]} \quad & \sum_{t=1}^{\tau} U(\min\{X[t], D[t]\}) - \delta \max_t \rho[t] X[t] \\ \text{s.t.} \quad & X[t] \geq 0, \qquad \forall t, \\ & \rho[t] \in \{0,1\}, \qquad \forall t, \\ & \sum_{t=1}^{\tau} \rho[t] = \lceil 0.95\tau \rceil. \end{aligned} \tag{5}$$

Here, $X[t]$ is the principal variable while $\rho[t]$ is the auxiliary variable that is used to calculate the $\mu_{95}(X[t])$. Note that, since the net utility function does not depend on the auxiliary variable $\rho[t]$, and also because price parameter δ is nonnegative, if the principal variable $X[t]$ is set to be fixed, then the maximization in (5) over $X[t]$ and $\rho[t]$ reduces to the minimization in (2) over $\rho[t]$. Therefore, it is guaranteed that once we solve the problem in (5), the choice of auxiliary variable $\rho[t]$ is automatically selected in a way that $\mu_{95}(X[t])$ is calculated as in (2). The solution of the deterministic problem (5) is the highest surplus that can be gained by the user, in case that the user's bandwidth demand can be perfectly predicted.

4.2 Optimization Problem with Stochastic Prediction

If the prediction of demand is uncertain, we maximize the user's surplus in an average sense, i.e., we maximize the user's expected surplus:

$$\sum_{t=1}^{\tau} \mathbb{E}\left\{U(\min\{X[t], D_k[t]\}) - \delta \cdot \mu_{95}(\bar{X}[t])\right\}, \tag{6}$$

A common approach in addressing uncertainty is to obtain a probability mass function [25] for each random parameter using historical workload data. This can be done in various levels of details and accuracy, e.g., see [26]. Specifically, we assume that each $D[t]$ has K_t possible realizations, $D_1[t], \ldots, D_{K_t}[t]$, where each realization $D_k[t]$ may occur with probability $\pi_{k,t}$. We have

$$\sum_{k=1}^{K_t} \pi_{k,t} = 1, \quad \forall t. \tag{7}$$

Once we use the above uncertainty modeling method, from (6), (7) and (2), we can formulate the following stochastic optimization problem to maximize the user's expected surplus over a billing cycle:

$$\begin{aligned} \max_{X[t],\rho[t]} \quad & \sum_{t=1}^{\tau} \sum_{k=1}^{K_t} \pi_{k,t} U(\min\{X[t], D_k[t]\}) - \delta \max_t \rho[t] X[t] \\ \text{s.t.} \quad & X[t] \geq 0, \qquad \forall t, \\ & \rho[t] \in \{0,1\}, \qquad \forall t, \\ & \sum_{t=1}^{\tau} \rho[t] = \lceil 0.95\tau \rceil. \end{aligned} \tag{8}$$

The only difference between problem (8) and (5) is the objective function. In problem (5), the CDC user makes a deterministic prediction of demand for bandwidth, i.e., parameters $D[1], \ldots, D[\tau]$, and thus the expectation of its utility gain at time slot t can be formulated by $U(\min\{X[t], D[t]\})$. While in problem (8), the CDC user makes a stochastic prediction, and thus the expectation of its utility gain at time slot t can be formulated by $\sum_{k=1}^{K_t} \pi_{k,t} U(\min\{X[t], D_k[t]\})$.

5 Solution Method

Both problems (5) and (8) are nonlinear mixed-integer programmings, which are generally considered to be hard problems to solve. Nevertheless, in this section, we explain how these problems can be solved with reasonable computational complexities.

5.1 Deterministic Problem Solution

For the deterministic problem (5), we can intuitively obtain its optimal solution for the auxiliary variables $\rho[1], \ldots, \rho[\tau]$ without numerically solving the problem. This property can be expressed mathematically in the following theorem.

Theorem 1. Let ϑ denote the set of all time slots t at which $D[t]$ is within the top 5% of the values in $D[1], \ldots, D[\tau]$. (a) There exists an optimal solution for the deterministic problem (5) in which the values of auxiliary variables $\rho[1], \ldots, \rho[\tau]$ are as follows:

$$\rho^\star[t] = \begin{cases} 0, & \forall t \in \vartheta; \\ 1, & \text{otherwise.} \end{cases} \tag{9}$$

(b) Once we replace the ρ in the deterministic problem (5) by (9), the optimal values of the principal variables $X[1], \ldots, X[\tau]$ of the deterministic problem (5) are obtained by solving the following convex optimization problem:

$$\begin{aligned} \max_{X[t]} \quad & \sum_{t=1}^{\tau} U(X[t]) - \delta \max_t \rho^\star[t] X[t] \\ \text{s.t.} \quad & 0 \leq X[t] \leq D[t], \ \forall t, \end{aligned} \tag{10}$$

where $\rho^\star[t]$ is given by (9).

The Proof of Theorem 1 is given in Appendix A. The theorem essentially implies that the optimal choice of usage *bursts* in (5) is the top 5% of the values in $D[1], \cdots, D[\tau]$. From Theorem 1, one can transform the non-convex problem (5) onto the convex program (10), which can be effectively solved using convex programming techniques [27].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

5.2 Stochastic Problem Solution

If parameters $D[1], \ldots, D[\tau]$ are random, then we do *not* know at what time slots the demand bursts will occur. Accordingly, we cannot use the approach discussed in Section 5.1 to figure out the optimal values of $\rho[1], \ldots, \rho[\tau]$. Therefore, we have no choice but solving the original stochastic problem (8).

A key difficulty in solving the stochastic problem (8) is that even if we relax the binary constraints, i.e., even if we choose $\rho[t]$ to be a continuous number between 0 and 1, the relaxed problem is still difficult to solve due to the non-convex term $\rho[t]X[t]$ in the objective function. Interestingly, we can tackle this undesirable property as it is explained in a theorem below.

Theorem 2. We can reformulate the stochastic problem (8) as

$$\begin{aligned}
\max_{X[t],\rho[t],\phi} \quad & \sum_{t=1}^{\tau}\sum_{k=1}^{K_t} \pi_{k,t} U(\min\{X[t], D_k[t]\}) - \delta \cdot \phi \\
\text{s.t.} \quad & X[t] \le \phi + L(1-\rho[t]), && \forall t, \\
& X[t] \ge 0, && \forall t, \\
& \rho[t] \in \{0,1\}, && \forall t, \\
& \sum_{t=1}^{\tau} \rho[t] = \lceil 0.95\tau \rceil,
\end{aligned} \tag{11}$$

where L is a large number compared to the available bandwidth, and ϕ is another auxiliary variable.

The proof of Theorem 2 is given in Appendix B. Given the equivalence of the stochastic problem (8) and (11), we can solve problem (11) instead of (8). Next, we notice that from (11), once we relax the binary constraints, the relaxed problem is convex. Therefore, we can find the exact optimal solution of problem (11) using branch-and-bound method [28], where at each branching step we need to solve a convex optimization problem. We refer to this approach as the convex branch-and-bound (CBB) method.

While the CBB method is effective to obtain the exact optimal solution of the stochastic surplus maximization problem, solving a nonlinear (although convex) problem at each iteration of the branch-and-bound algorithm could be time consuming. Since the nonlinearity in problem (11) is due to the nonlinearity of the utility function $U(\cdot)$, one way to make problem (11) linear is to replace $U(\cdot)$ with its piecewise linear approximation:

$$\begin{aligned}
\max_{\substack{X[t],\rho[t],\phi,\\ Q_k[t],h_k[t]}} \quad & \sum_{t=1}^{\tau}\sum_{k=1}^{K_t} \pi_{k,t} h_k[t] - \delta \cdot \phi \\
\text{s.t.} \quad & X[t] \le \phi + L(1-\rho[t]), && \forall t, \\
& X[t] \ge 0, && \forall t, \\
& \rho[t] \in \{0,1\}, && \forall t, \\
& \sum_{t=1}^{\tau} \rho[t] = \lceil 0.95\tau \rceil, \\
& Q_k[t] \le X[t], && \forall t, k, \\
& Q_k[t] \le D_k[t], && \forall t, k, \\
& h_k[t] \le U(n\Delta[t]) + \\
& \qquad U'(n\Delta[t])(Q_k[t] - n\Delta[t]), && \forall t, k, n,
\end{aligned} \tag{12}$$

where $n = 1, \ldots, N$, and N is the number of segments in piecewise linearizing the utility function U. Also, $h_k[t]$ and $Q_k[t]$ are auxiliary variables. Since problem (12) maximizes $h_k[t]$, and $h_k[t]$ appears only in the last constraint, at optimality we have $Q_k[t] = \max\{X[t], D_k[t]\}$. Consequently, from the last constraint in (12) at optimality the variable$h_k[t]$ is the piecewise linearized $U(\max\{X[t], D_k[t]\})$ at the points $n\Delta$. Therefore, problem (12) is equivalent to problem (11) when $N \to \infty$. The solution of problem (12) depends on the choice of parameter N. However, as we will discuss further, the problem (12) gives a solution near optimal solution of problem (11) even for smaller N. Problem (12) is a mixed-integer linear program (MILP) and can be solved by existing solvers such as CPLEX [29]. In Section 7.2 we will see that the computation time of solving problem (12) is substantially less that the computation time of the CBB method.

Before we end this section, we must point out that one can obtain an *approximate* solution for problem (12) by terminating the optimization solver at certain guaranteed optimality bounds in order to significantly lower computational complexity. We will further discuss this option in Section 7.2.

6 Extensions and Remarks

We may extend our design to a scenario where a user has the option to receive service from multiple providers. An example for this scenario is when a user can download specified content over different transit links that is owned by different ISPs, who charge the user via burstable billing. In this section, we also show that a user can further improve its surplus by updating the usage of bandwidth in real-time, i.e. during the billing cycle, based on the newly *exposed* actual demand information.

6.1 Extension to Multiple Providers

Let $X_i[t]$ denote the *planned* usage of bandwidth at provider i at time slot t decided based on the demand $D[t]$. Let δ_i (\$/Mbps) denote the price of bandwidth at provider i. In

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

this case, in each billing cycle, the *expected* surplus of the user is obtained as

$$S = \sum_{t=1}^{\tau} \mathbb{E}\left\{U(\min\{\sum_{i=1}^{I} X_i[t], D[t]\})\right\} - \sum_{i=1}^{I} \delta_i \cdot \mu_{95}(X_i[t]). \tag{13}$$

From (7) and (13), following optimization problem maximizes user's expected surplus:

$$\begin{aligned} \max_{X_i[t],\rho_i[t]} \quad & \sum_{t=1}^{\tau}\sum_{k=1}^{K_t} \pi_{k,t} U(\min\{\sum_{i=1}^{I} X_i[t], D_k[t]\}) - \\ & \sum_{i=1}^{I} \delta_i \max_t \rho_i[t] X_i[t] \\ \text{s.t.} \quad & X_i[t] \geq 0, && \forall t, i, \\ & \rho_i[t] \in \{0,1\}, && \forall t, i, \\ & \sum_{t=1}^{\tau} \rho_i[t] = \lceil 0.95\tau \rceil, && \forall i. \end{aligned} \tag{14}$$

The deterministic user's surplus maximization is a special case of problem (14) where $\forall t$, $K_t = 1$ and $\pi_{k,t} = 1$. Note that, for the case of multiple providers, the exact solution of the optimization in (14), even for the deterministic optimization, cannot be obtained from the method discussed in Theorem 1. Therefore, we propose the following approach for solution of the problem in (14).

As in Section 5.2, we can transform the nonlinear mixed-integer programming (14) into an equivalent mixed-integer convex programming (11) or MILP (12) , where the mixed-integer convex programming and the MILP can be solved via CBB and MILP solvers such as CPLEX [29], respectively.

6.2 Updating Usage of Bandwidth During a Cycle

Next, we show that the user can further improve its surplus during a billing cycle, by updating its *planned* usage of bandwidth at each time slot based on the newly *exposed* actual demand. Therefore the user's *final* surplus *after* a cycle will be no less than expected surplus. Here, we assume that, at the beginning of each time slot, the user's demand for bandwidth is *exposed* to the user. We denote the *exposed* demand value at time slot t by $\bar{D}[t]$.

Generally, the demand $D[t]$ may not be the same as the *exposed* value $\bar{D}[t]$. Therefore, a user can *update* its *planned* usage of bandwidth in real-time based on the newly learned *exposed* demand information, i.e., $\bar{D}[t]$, to further improve its surplus while keeping its bandwidth cost unchanged. For example, if $X[t] < \bar{D}[t]$ and $X[t] < \mu_{95}(X[t])$, the user can increase its usage from $X[t]$ to $\min\{\bar{D}[t], \mu_{95}(X[t])\}$. In this way, the user's net utility can be enhanced while remaining its bandwidth cost unchanged.

In practice, the *expected 95th percentile usage* $\mu_{95}(X[t])$ is treated as a rate limiter. According to (2), when $\rho[t] = 1$, the user restricts its usage at this times slot to reduce its *95th percentile usage*. Specifically, when $\rho[t] = 1$, if $\bar{D}[t] \leq \mu_{95}(X[t])$, the user can utilize bandwidth on-demand, and if $\bar{D}[t] > \mu_{95}(X[t])$, the user needs to restrict its utilization of bandwidth to ensure that its *95th percentile usage* equals to $\mu_{95}(X[t])$. On the contrary, the user can always utilize bandwidth on-demand when $\rho[t] = 0$ since the usage at this time slot has no impact on the *95th percentile usage*. Therefore, we formulate the user's *updated* usage of bandwidth at each time slot, which is denoted by $\bar{X}[t]$, as

$$\bar{X}[t] = \begin{cases} \bar{D}[t], & \text{if } \rho[t] = 0 \text{ or } \bar{D}[t] \leq \mu_{95}(X[t]); \\ \mu_{95}(X[t]), & \text{otherwise.} \end{cases} \tag{15}$$

From (15), we ensure that $\forall t$, $\bar{X}[t] \leq \bar{D}[t]$. Similar to (3) and (4), after a billing cycle, the net utility with *updated* usage values $\bar{X}[1], \ldots, \bar{X}[\tau]$ can be calculated as

$$\bar{R} = \sum_{t=1}^{\tau} U(\bar{X}[t]). \tag{16}$$

Further, from (2), (1) and (16), we formulate the user's surplus with *updated* usage values $\bar{X}[1], \ldots, \bar{X}[\tau]$ via

$$\bar{S} = \sum_{t=1}^{\tau} U(\bar{X}[t]) - \delta \cdot \mu_{95}(\bar{X}[t]). \tag{17}$$

We can show that a user's surplus with *updated* usage values $\bar{X}[1], \ldots, \bar{X}[\tau]$ is always no less than its surplus with *planned* usage values $X[1], \ldots, X[\tau]$. From (15), we ensure that $\mu_{95}(\bar{X}[t]) \leq \mu_{95}(X[t])$. Therefore, the bandwidth cost over a billing cycle with *updated* usage values $\bar{X}[t]$ is always no higher than the bandwidth cost with *planned* usage values $X[t]$.

Next, we notice that the net utility over a billing cycle with *updated* usage values $\bar{X}[t]$ is always no less than the bandwidth cost with *planned* usage values $X[t]$, i.e.,

$$U(\min\{\bar{X}[t], \bar{D}[t]\}) \geq U(\min\{X[t], \bar{D}[t]\}), \quad \forall t. \tag{18}$$

To verify that (18) indeed holds, consider three cases:
Case 1: If $\rho[t] = 0$, $\bar{X}[t] = \bar{D}[t]$. Since the net utility function $U(\cdot)$ is nondecreasing and $T > 0$, (18) is satisfied.
Case 2: If $\rho[t] = 1$ and $\bar{D}[t] \leq \mu_{95}(X[t])$, $\bar{X}[t] = \bar{D}[t]$. Same as case 1, in this case, (18) is satisfied.
Case 3: If $\rho[t] = 1$ and $\bar{D}[t] > \mu_{95}(X[t])$, $\bar{X}[t] = \mu_{95}(X[t])$ and $X[t] \leq \mu_{95}(X[t])$. In this case, (18) is also satisfied.

Accordingly, it can readily be concluded that a user's surplus with *updated* usage values is also no less than its surplus with *planned* usage values.

Identically, if the user can receive service from multiple providers, we can also update its *planned* usage of bandwidth at provider i, i.e., $X[t]$, in real-time based on the newly learned information of the *exposed* demand $\bar{D}[t]$. Let $\bar{X}_i[t]$ denote the *updated* usage of bandwidth at provider i at time slot t and it is defined as

$$\bar{X}_i[t] = \begin{cases} \bar{D}[t], & \text{if } \rho_i[t] = 0 \text{ or } \bar{D}[t] \leq \mu_{95}(X_i[t]); \\ \mu_{95}(X_i[t]), & \text{otherwise,} \end{cases} \tag{19}$$

where $\rho_i[t]$ is the auxiliary variable as used in (2). Then, we formulate the user's surplus over a billing cycle via

$$\bar{S} = \sum_{t=1}^{\tau} U(\sum_{i=1}^{I} \bar{X}_i[t]) - \sum_{i=1}^{I} \delta_i \cdot \mu_{95}(\bar{X}_i[t]). \tag{20}$$

Similarly, a user can also further improve its surplus via updating its *planned* usage according to (19) if it can receive service from multiple providers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

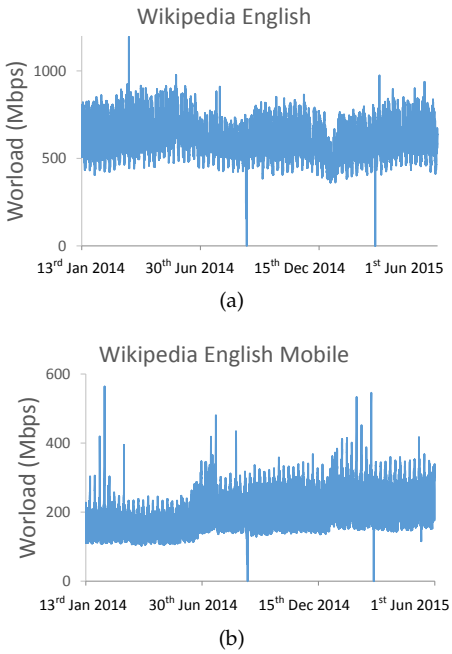


Fig. 4. Examples for the real-world workload traces used in this paper from [12]; a) data trace of Wikipedia English, b) data trace of Wikipedia English Mobile.

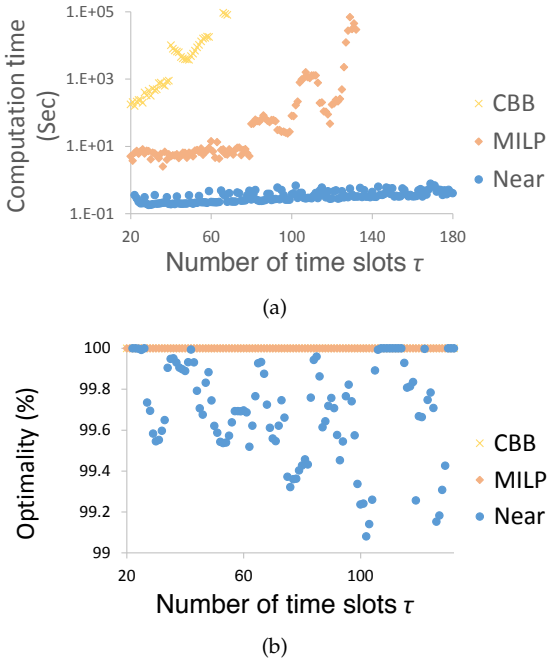


Fig. 5. Comparing different solution methods in solving problem (11): (a) Computation time, (b) Optimality.

Note that, since the final surplus a user can achieve in our design is obtained from (17) and (20), we use these values as the user's surplus, in the rest of this paper.

7 Case Studies

In this section, with real-world data traces, we first study the computation time and performance of our proposed solution methods for solving the stochastic problem (11). Second, we evaluate the performance of our design with a simple method to forecast the demand for bandwidth. Third, we discuss the impact of price and utility factor on the performance of our design. Forth, we show that, in the presence of multiple providers, the user can further improve its surplus by using our design.

7.1 Setup

We use two data sets in our case studies: 1) *Wikien*: the page view data of Wikipedia English from January 2014 to May 2015 [12], 2) *Wikimw*: the page view data of Wikipedia English Mobile from January 2014 to May 2015 [12]. Example traces of these data sets are shown in Fig. 4. Each time slot takes one hour and the billing cycle takes 28 days for Wikien and Wikimw data sets.

The utility functions are selected as follows [30], [31]:

$$U(x) = \begin{cases} A(1-a)^{-1}x^{1-a}, & \text{if } a \in (0,1); \\ Alog(x), & \text{if } a = 1. \end{cases} \quad (21)$$

Here, $A > 0$ is the utility factor decided by the user and $a \in (0,1]$ measures the concavity of the user's utility. Namely, as a increases, the user's utility becomes more concave. Specifically, we assume that $a = 0.1$, $A = 0.08$ and the impact of the utility factor A on the surplus of user will be discussed in Section 7.4.

We use a very simple workload forecasting method. Let $D_1[t]$ and $D_2[t]$ denote the workload at time slot t in the last two billing cycles, respectively. Suppose that $\pi_{1,t} = \pi_{2,t} = 0.5$, $\forall t = 1, \ldots, \tau$. Specifically, for deterministic surplus maximization, we assume that $D[t] = \pi_{1,t}D_1[t] + \pi_{2,t}D_2[t]$, for any $t = 1, \ldots, \tau$.

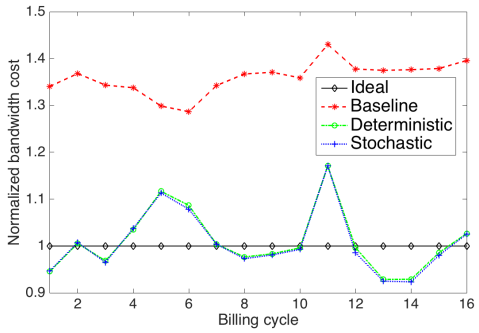
7.2 Computation Complexity of Proposed Solution Methods

Recall from Section 5.2 that there are multiple options to solve the stochastic problem (11). Specifically, the proposed CBB method leads to the exact optimal solution. The efficiency of the MILP method, however, depends on the number of tangent lines N. Here, we let $\Delta[t] = TD_k[t]/N$ and we assume that $N = 3$.

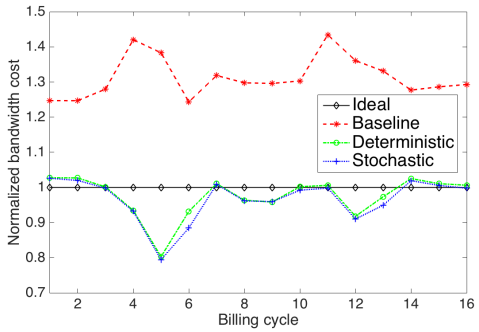
We evaluate the computation time for each solution method. We use a personal computer with Intel Xeon CPU E5-2450 @2.50GHZ. The results are shown in Fig. 5(a). We can see that the computation time of CBB is much longer than MILP. Even for the MILP approach, it may take several hours to find the global optimal solution of problem (12) as the size of the problem increases.

As we pointed out in Section 5.2, one can obtain an *approximate* solution for problem (12) by terminating the optimization solver at certain guaranteed optimality bounds. This can be done by setting up a stopping condition for the MILP method based on the ratio between the upper-bound and the lower-bound solutions. The upper-bound solution is the surplus that can be achieved if we relax the remaining binary variables at the current branching stage. The lower-bound solution is the surplus at the best binary solution that has been obtained so far at the current branching stage. Clearly, this ratio indicates a guaranteed optimality in the solution of MILP that has already been reached at the current branching stage. In this paper, we obtain an approxi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

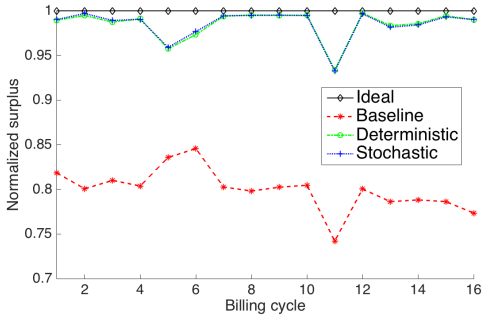


(a)

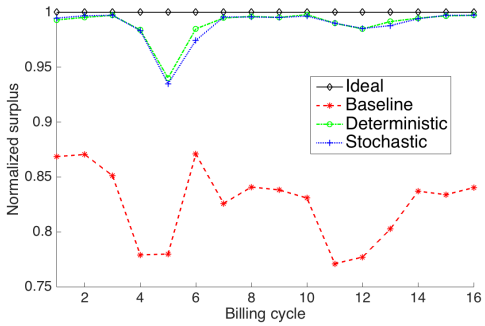


(b)

Fig. 6. Comparing normalized bandwidth cost under different methods and different workloads: a) Wikien, b) Wikimw.



(a)



(b)

Fig. 7. Comparing normalized surplus under different methods and different workloads: a) Wikien, b) Wikimw.

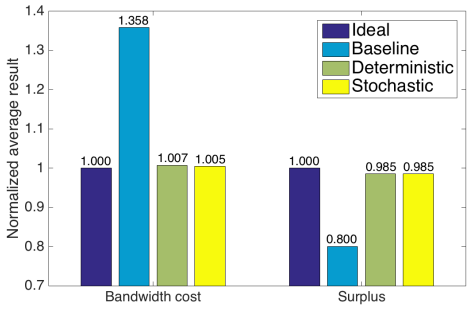
mate solution by stopping the MILP method in CPLEX once the above mentioned ratio reaches 5%, which guarantees at least 95% optimality. We refer to this approximate solution approach as the *Near* method.

Fig. 5 shows the comparison among CBB, MILP and Near in computation time and result of optimality. As we can see in Fig. 5(a), the Near method is significantly less complex in terms of required computation, compared to the CBB and MILP methods. Specifically, the computational time for the Near method grows only linearly with respect to the number of time slots while CBB and MILP grow exponentially. Interestingly, we can see in Fig. 5(b) that the actual achieved optimality is around 99% or more, i.e., much better than the guaranteed 95% worst case optimality value. Therefore, for the rest of this paper, we use the Near method at 95% guaranteed optimality.

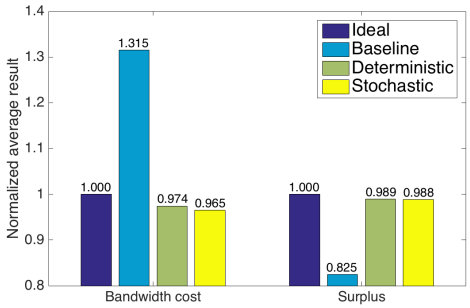
7.3 Performance Evaluation

As a *Baseline* for performance comparison, we consider the case where the bandwidth is allocated on-demand, i.e., $X[t] = \bar{X}[t] = \bar{D}[t]$, for any $t = 1, \ldots, \tau$. Note that, this approach resembles how the bandwidth is currently allocated in practice. Next, we also assume an *Ideal* case where the usage of bandwidth is optimized based on *true knowledge* of demand, i.e., $\forall t$, $D[t] = \bar{D}[t]$. While the Baseline shows how well we can perform compared to the existing practice, the Ideal case shows the best performance that we can ever get, assuming that we can perfectly predict the upcoming workload.

Next, we compare the Baseline and Ideal cases with our proposed *Deterministic* and *Stochastic* methods. The Deterministic method refers to the case where the bandwidth usage is scheduled based on the optimal solution of the deterministic surplus maximization problem in (10). The Stochastic method refers to the case where the bandwidth usage is scheduled based on the optimal solution of the



(a)



(b)

Fig. 8. Comparing average bandwidth cost and surplus under different methods and different workloads: a) Wikien, b) Wikimw.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

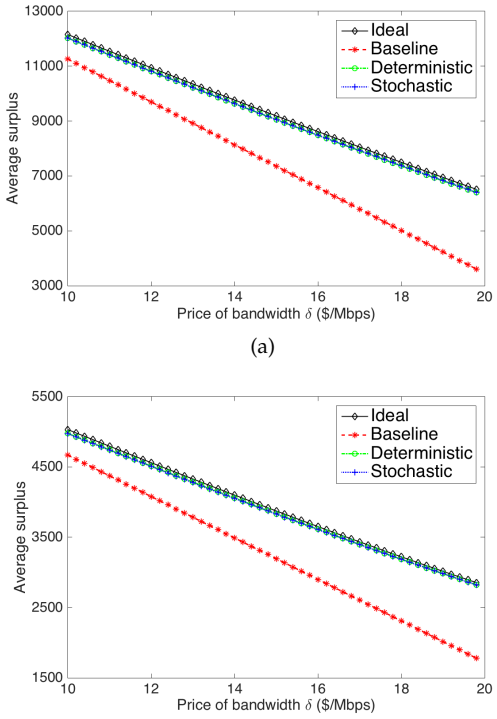


Fig. 9. The impact of the price of bandwidth on average surplus under different workloads: a) Wikien, b) Wikimw.

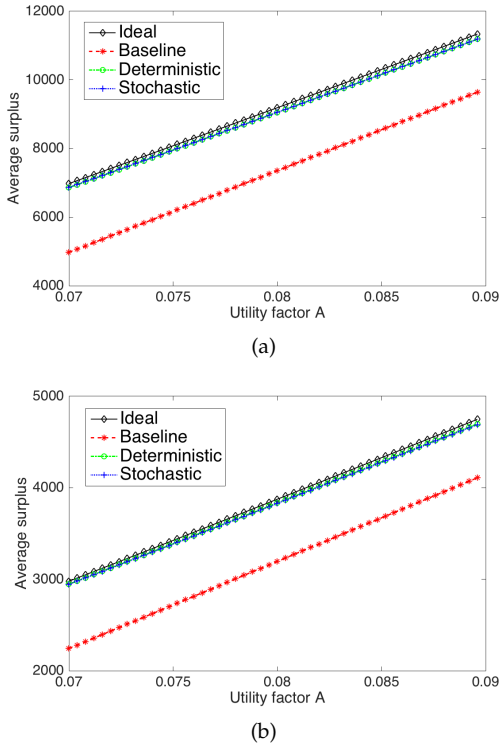


Fig. 10. The impact of the utility factor on average surplus under different workloads: a) Wikien, b) Wikimw.

stochastic surplus maximization problem in (12) using the Near method with 95% guaranteed optimality. The method of forecasting the workload in each case was already explained in Section 7.1.

The results on performance comparison are shown in Fig. 6, Fig. 7 and Fig. 8, where the results for all methods are normalized with respect to the results of the Ideal case. Here, the price of bandwidth is set to be $15 per Mbps. We can make the following observations based on these results:

- As shown in Fig. 6 and Fig. 8, Deterministic and Stochastic solutions may result in less bandwidth cost than Ideal, due to under-prediction of demands. In this case, their bandwidth costs reduce, as well as their surpluses.
- As shown in Fig. 6 and Fig. 7, even though we use a very simple method to forecast the demand for bandwidth, the Deterministic and Stochastic solutions outperform the Baseline in both bandwidth cost reduction and surplus improvement. Meanwhile, Deterministic and Stochastic have similar outcomes. Thus, our method is robust to the error of prediction of user's demand.
- As shown in Fig. 8, on average, our proposed optimization-based approach to respond to burstable billing can greatly reduce the user's bandwidth cost while improving its surplus when comparing against Baseline. For example, with data trace of Wikien, both Deterministic and Stochastic surplus maximization can reduce the user's bandwidth cost by 26% while increasing its total surplus by 23%.

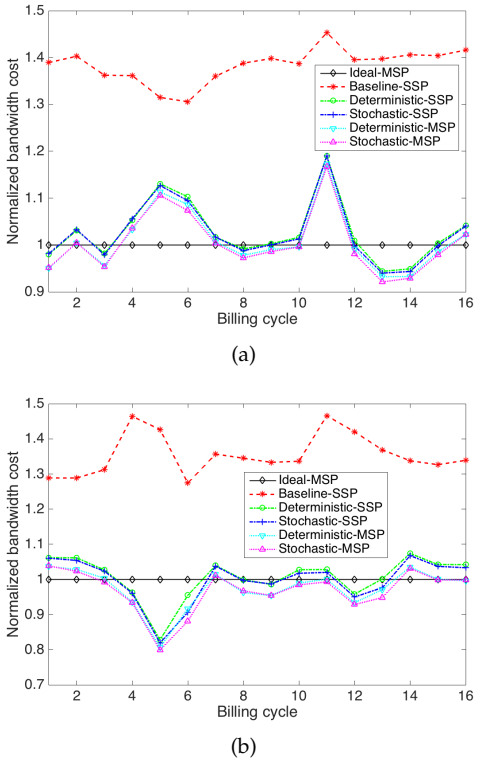


Fig. 11. Comparing normalized bandwidth cost with multiple providers under different workloads: a) Wikien, b) Wikimw.

7.4 Impact of Price and Utility Factor

Intuitively, increasing the price for bandwidth would increase the user's cost. Accordingly, the surplus that the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

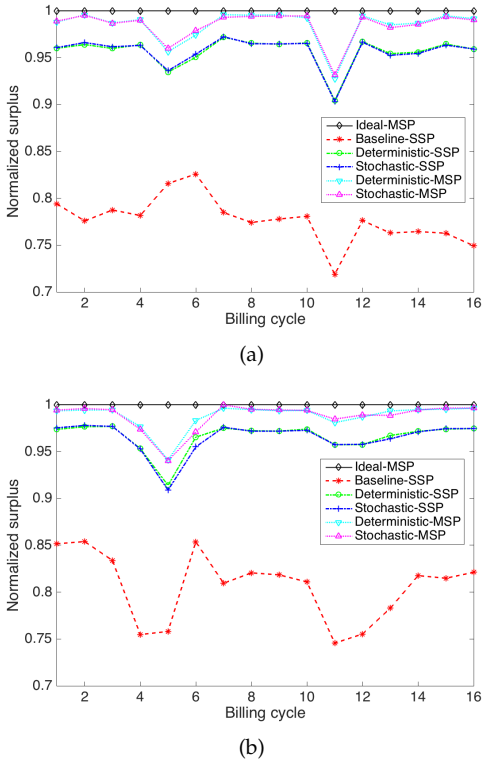


Fig. 12. Comparing normalized surplus with multiple providers under different workloads: a) Wikien, b) Wikimw.

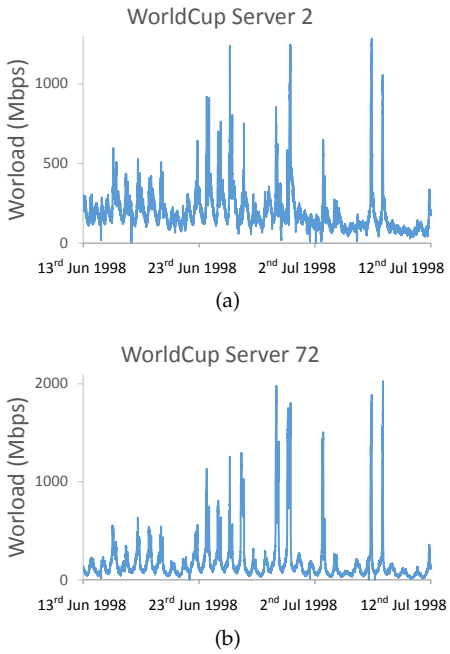


Fig. 13. Examples for the real-world workload traces used in this paper from [32]; a) web hits to the 2nd server during WorldCup 98, b) web hits to the 72nd server during WorldCup 98.

user may gain decreases as we increase price parameter δ. However, the rate of such decrease is *not* the same for different methods. The results are shown in Fig. 9. We can see that the rate of decrease in surplus is higher for the Baseline compared to the Deterministic and Stochastic methods. As a results, the surplus improvements with our proposed optimization-based approaches are higher when the price of bandwidth is high.

Next, we analyze the impact of utility factor A. Clearly, increasing the parameter A in (21) results in higher surplus for the same usage of bandwidth. By analysing Fig. 10, we find that the distance between Baseline and Deterministic/Stochastic is slightly larger when parameter A is small. Namely, users with smaller utility factors, who are more sensitive to price than performance, are more likely to response to the burstable billing to improve their surpluses. We can also see that the Deterministic and Stochastic methods outperform the Baseline at all choices of parameter A.

7.5 Impact of Multiple Providers

Suppose the user can receive service from two providers, who are referred to as providers 1 and 2. Both of them offer bandwidth at \$15 per Mbps. To evaluate our proposed approach to response to burstable billing with multiple providers, we simulate six different cases:

- *Ideal-MSP*: It is defined as the outcome of maximizing surplus, under the assumption that the demand for bandwidth is known with multiple providers.
- *Baseline-SSP*: In this case, the user utilizes bandwidth from provider 1 on-demand.
- *Deterministic-SSP*: In this case, the user utilizes bandwidth from provider 1 and makes its decisions based on our design with deterministic prediction about its demand.
- *Stochastic-SSP*: In this case, the user utilizes bandwidth from provider 1 and makes its decisions based on our design with stochastic prediction about its demand.
- *Deterministic-MSP*: In this case, the user utilizes bandwidth from both provider 1 and 2 and makes its decisions based on our design with deterministic prediction about its demand.
- *Stochastic-MSP*: In this case, the user utilizes bandwidth from both provider 1 and 2 and makes its decisions based on our design with stochastic prediction about its demand.

Figures 11 and 12 show the normalized bandwidth cost and surplus, obtained in six different cases, where the base for normalization is the surplus under the Ideal-MSP case. We can see that Deterministic-MSP and Stochastic-MSP methods always outperform Baseline-SSP in both bandwidth cost reduction and surplus improvement. Finally, we also find that Deterministic-MSP and Stochastic-MSP are always better than Deterministic-SSP and Stochastic-SSP. We may infer that the availability of multiple providers further reduce the user's bandwidth cost and improves its surplus under optimal response mechanism to burstable billing.

7.6 Performance Evaluation with "Bursty" Data Traces

In this section we further evaluate our proposed mechanism using additional data traces:

- *WC2*: 0.1 percent of web hits to the 2nd server during WorldCup 98 from June 13, 1998 to July 12, 1998 [32];

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

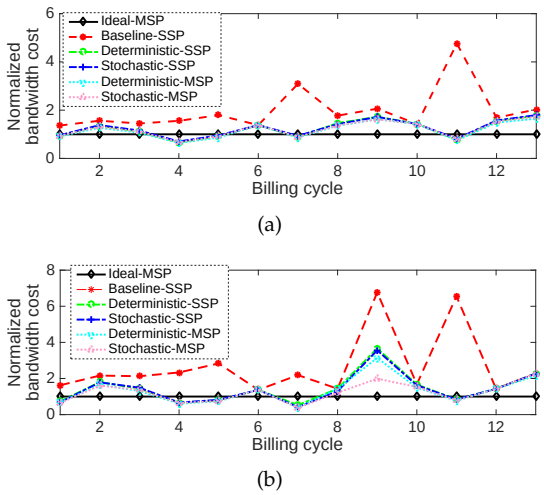


Fig. 14. Comparing normalized bandwidth cost with multiple providers under different workloads: a) WC2, b) WC72.

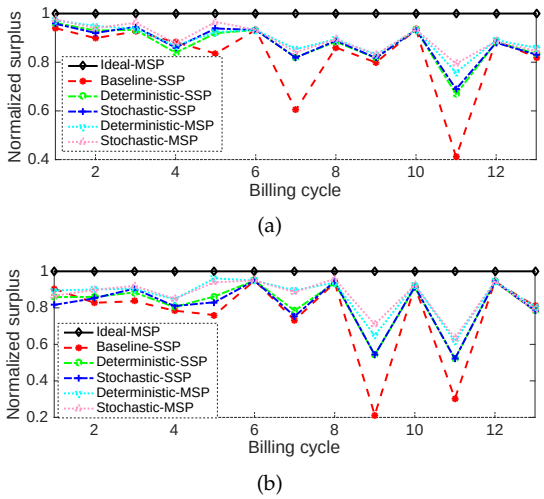


Fig. 15. Comparing normalized surplus with multiple providers under different workloads: a) WC2, b) WC72.

- *WC72*: 0.1 percent of web hits to the 72nd server during WorldCup 98 from June 13, 1998 to July 12, 1998 [32].

Example traces of these data sets are shown in Fig. 13. Comparing Fig. 13 with Fig. 4, one can easily find that the WorldCup data traces as shown in Fig. 13 is quite "bursty".

Figures 14 and 15 show the normalized bandwidth cost and surplus with WorldCup data traces, where the base for normalization is the surplus under the Ideal-MSP case. We can see that Deterministic-MSP and Stochastic-MSP methods outperform Baseline-SSP in most cases.

8 Conclusion and Future Work

A novel cost-aware traffic management mechanism was proposed to select the usage of bandwidth for a CDC user, who is charged for bandwidth usage under burstable billing. Our proposed mechanism considers workload demand uncertainty, and is general in the sense that it does not make any assumption about the statistical characteristics of workload. Numerical results based on empirical case studies confirm that even with a simply workload forecasting method, the user can significantly reduce its IP transit cost while increasing its surplus with our proposed mechanism for responding to burstable billing, compared to the current practice of allocating bandwidth on-demand. We also extended our design to another emerging practical scenario where a user can receive service from multiple providers. Accordingly, besides bandwidth allocation, our problem formulation also addresses workload distribution.

This paper can be extended in several directions. First, one can adopt a more advanced workload forecasting method to better model probability distribution functions for the demand for bandwidth. In fact, with enough accurate prediction, the performance of the proposed methods are guaranteed to improve. Second, one can try to further reduce a user's *95th percentile usage* via traffic shaping [21], traffic aggregation [7], traffic shifting in time and space [13], simultaneously. Finally, one can revisit the problem from the provider's viewpoint based on the knowledge of how a user optimally responds to burstable billing and adjusts the billing parameters to achieve better results for the provider.

Appendix A
Proof of Theorem 1

Let $\bar{\vartheta}$ denote the complement set of ϑ, i.e., $\bar{\vartheta} = \{1, \cdots, \tau\} - \vartheta$. Problem (5) is always feasible and therefore has at least one solution $(\rho^*[t], X^*[t])$. If $\exists t_\vartheta \in \vartheta$ such that $\rho[\vartheta] = 1$, then from the last constraint in problem (5) $\exists t_{\bar{\vartheta}} \in \bar{\vartheta}$ for which $\rho[t_{\bar{\vartheta}}] = 0$. In order to prove Theorem 1, we only have to show that there exist an optimal solution for problem (5), for which

$$\tilde{\rho}[t] = \begin{cases} 0 & t = t_\vartheta \\ 1 & t = t_{\bar{\vartheta}} \\ \rho^*[t] & \text{Otherwise.} \end{cases} \tag{22}$$

First we notice that, problem (5) implies that

$$\tilde{X}[t_\vartheta] = D[t_\vartheta] \quad \text{and} \quad X^*[t_{\bar{\vartheta}}] = D[t_{\bar{\vartheta}}]. \tag{23}$$

We Consider two cases:

Case 1, where:

$$D[t_{\bar{\vartheta}}] \geq X^*[t_\vartheta]. \tag{24}$$

Let $\tilde{X}[t]$ denote

$$\tilde{X}[t] = \begin{cases} D[t_\vartheta] & t = t_\vartheta \\ X^*[t_\vartheta] & t = t_{\bar{\vartheta}} \\ X^*[t] & \text{Otherwise.} \end{cases} \tag{25}$$

The couple $(\tilde{\rho}[t], \tilde{X}[t])$ is feasible in problem (5) and we have:

$$\max_t \tilde{\rho}[t]\tilde{X}[t] = \max_t \rho^*[t]X^*[t]. \tag{26}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2018.2796095, IEEE Transactions on Services Computing

Also,

$$\sum_{t=1}^{\tau} U(\min\{\tilde{X}[t], D[t]\}) - \sum_{t=1}^{\tau} U(\min\{X^*[t], D[t]\}) = \big(U(\min\{\tilde{X}[t_{\bar{\vartheta}}], D[t_{\bar{\vartheta}}]\}) + U(\min\{\tilde{X}[t_{\vartheta}], D[t_{\vartheta}]\})\big) - \big(U(\min\{X^*[t_{\bar{\vartheta}}], D[t_{\bar{\vartheta}}]\}) + U(\min\{X^*[t_{\vartheta}], D[t_{\vartheta}]\})\big) = (U(X^*[t_{\vartheta}]) + U(D[t_{\vartheta}])) - (U(D[t_{\bar{\vartheta}}]) + U(\min\{X^*[t_{\vartheta}], D[t_{\vartheta}]\})) = (U(D[t_{\vartheta}]) - U(D[t_{\bar{\vartheta}}])) + (U(X^*[t_{\vartheta}]) - U(\min\{X^*[t_{\vartheta}], D[t_{\vartheta}]\})) \geq 0. \tag{27}$$

The first equality is concluded from (22) and (25), and the second equality is concluded from (23) and (24). Also, the inequality is concluded from the fact that:

$$D[t] \geq D[t'] \quad \forall t \in \vartheta,\ \forall t' \in \bar{\vartheta}, \tag{28}$$

and

$$U(X^*[t_{\vartheta}]) \geq U(\min\{X^*[t_{\vartheta}], D[t_{\vartheta}]\}). \tag{29}$$

From (26) and (27), $(\tilde{\rho}[t], \tilde{X}[t])$ gives an objective value for problem (5) that is no less that the objective value that $(\rho^*[t], X^*[t])$ gives for the same problem. Therefore, $(\tilde{\rho}[t], \tilde{X}[t])$ is an optimal solution for problem (5).

Case 2, where:

$$D[t_{\bar{\vartheta}}] < X^*[t_{\vartheta}]. \tag{30}$$

Let $\tilde{X}[t]$ denote

$$\tilde{X}[t] = \begin{cases} D[t_{\vartheta}] & t = t_{\vartheta} \\ D[t_{\bar{\vartheta}}] & t = t_{\bar{\vartheta}} \\ X^*[t] & \text{Otherwise.} \end{cases} \tag{31}$$

The couple $(\tilde{\rho}[t], \tilde{X}[t])$ is feasible in problem (5) and we have:

$$\max_t \tilde{\rho}[t]\tilde{X}[t] \leq \max_t \rho^*[t]X^*[t], \tag{32}$$

where the inequality is concluded from (30). Also,

$$\sum_{t=1}^{\tau} U(\min\{\tilde{X}[t], D[t]\}) - \sum_{t=1}^{\tau} U(\min\{X^*[t], D[t]\}) = \big(U(\min\{\tilde{X}[t_{\bar{\vartheta}}], D[t_{\bar{\vartheta}}]\}) + U(\min\{\tilde{X}[t_{\vartheta}], D[t_{\vartheta}]\})\big) - \big(U(\min\{X^*[t_{\bar{\vartheta}}], D[t_{\bar{\vartheta}}]\}) + U(\min\{X^*[t_{\vartheta}], D[t_{\vartheta}]\})\big) = (U(D[t_{\bar{\vartheta}}]) + U(D[t_{\vartheta}])) - (U(D[t_{\bar{\vartheta}}]) + U(\min\{X^*[t_{\vartheta}], D[t_{\vartheta}]\})) \geq 0, \tag{33}$$

where the first equality is concluded from (30), and the inequality concluded from the fact that:

$$D[t_{\vartheta}] \geq U(\min\{X^*[t_{\vartheta}], D[t_{\vartheta}]\})). \tag{34}$$

From (32) and (33), $(\tilde{\rho}[t], \tilde{X}[t])$ gives an objective value for problem (5) that is no less that the objective value that $(\rho^*[t], X^*[t])$ gives for the same problem. Therefore, $(\tilde{\rho}[t], \tilde{X}[t])$ is an optimal solution for problem (5).

APPENDIX B
PROOF OF THEOREM 2

At each time slot t, if $\rho[t] = 0$, then the first constraint in problem (11) reduces to $X[t] \leq \phi + L,\ \forall t$, which always holds regardless of the values of $X[t]$ and ϕ. If $\rho[t] = 1$, then the first constraint in (11) reduces to $X[t] \leq \phi,\ \forall t$. In that case, since the objective function in (11) is to minimize ϕ, we necessarily obtain that $\phi = \max_t \rho[t]X[t]$ at any optimal solution of problem (11). This is clearly an outcome that we intended.

REFERENCES

[1] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2015–2020,", 2016.

[2] A. Odlyzko, "Internet pricing and the history of communications," *Computer Networks*, vol. 36, pp. 493–517, Aug 2001.

[3] X. Dimitropoulos, P. Hurley, A. Kind, and M. P. Stoecklin, "On the 95-percentile billing method," *Passive and Active Network Measurement*, vol. 5448, pp. 207–216, 2009.

[4] V. Reddyvari Raja, A. Dhamdhere, A. Scicchitano, S. Shakkottai, k. claffy, and S. Leinen, "Volume-based transit pricing: Is 95 the right percentile?," *Lecture Notes in Computer Science*, vol. 8362, pp. 77–87, 2014.

[5] A. Sathiaseelan, G. Tyson, and S. Sen, "Exploring the role of smart data pricing in enabling affordable internet access," in *Proc. of IEEE INFOCOM WKSHPS*, Hong Kong, China, Apr 2015.

[6] V. R. Raja, S. Shakkottai, A. Dhamdhere, and k. claffy, "Fair, flexible and feasible isp billing," *SIGMETRICS Perform. Eval. Rev.*, vol. 42, pp. 25–28, Dec 2014.

[7] I. Castro, R. Stanojevic, and S. Gorinsky, "Using tuangou to reduce ip transit costs," in *IEEE/ACM Trans. on Networking*, vol. 22, pp. 1415–1428, Oct 2014.

[8] J. Garciadorado and S. Rao, "Cost-aware Multi Data-Center Bulk Transfers in the Cloud from a Customer-Side Perspective," *IEEE Transactions on Cloud Computing*, vol. PP, pp. 1–1, 2015.

[9] Y. Chen, S. Jain, V. K. Adhikari, Z. li Zhang, and K. Xu, "A first look at inter-data center traffic characteristics via Yahoo! datasets," in *Proc. of IEEE INFOCOM*, Shang Hai, China, Apr 2011.

[10] E. Zohar, I. Cidon, and O. Mokryn, "The power of prediction: cloud bandwidth and cost reduction," in *Proc. of ACM SIGCOMM*, Toronto, Canada, Aug 2011.

[11] A. Gandhi, C. Yuan, D. Gmach, M. Arlitt and M. Marwah, "Minimizing Data Center SLA Violations and Power Consumption via Hybrid Resource Provisioning," in *Proc. of IEEE IGCC*, Orlando, USA, Jul 2011.

[12] "Page view statistics for Wikimedia projects", http://dumps.wikimedia.org/other/pagecounts-raw/.

[13] R. G. Clegg, R. Landa, J. T. Arajo, E. Mykoniati, D. Griffin, and M. Rio, "Tardis: Stably shifting traffic in space and time," *SIGMETRICS Perform. Eval. Rev.*, vol. 42, pp. 593–594, Jun 2014.

[14] L. Zhang, Z. Li, C. Wu, and M. Chen, "Online algorithms for uploading deferrable big data to the cloud," in *Proc. of IEEE INFOCOM*, Toronto, ON, Apr 2014.

[15] X. Xiang, C. Lin, F. Chen, and X. Chen, "Greening geo-distributed data centers by joint optimization of request routing and virtual machine scheduling," in *Proc. of IEEE/ACM UCC*, London, UK, Dec 2014.

[16] H. Xu and B. Li, "Joint request mapping and response routing for geo-distributed cloud services," in *Proc. of IEEE INFOCOM*, Turin, Italy, Apr 2013.

[17] H. Xu and B. Li, "Cost efficient datacenter selection for cloud services," in *Proc. of IEEE/CIC ICCC*, Beijing, China, Aug 2012.

[18] S. Traverso, K. Huguenin, I. Trestian, V. Erramilli, N. Laoutaris, and K. Papagiannaki, "Social-aware replication in geo-diverse online systems," *IEEE Trans. on Parallel and Distributed Systems*, vol. 26, pp. 584–593, Feb 2015.

[19] L. Golubchik, S. Khuller, K. Mukherjee, and Y. Yao, "To send or not to send: Reducing the cost of data transmission," in *Proc. of IEEE INFOCOM*, Turin, Italy, Apr 2013.

[20] T. Nandagopal and K. P. Puttaswamy, "Lowering inter-datacenter bandwidth costs via bulk data scheduling," in *Proc. of IEEE/ACM CCGrid*, Ottawa, ON, May 2012.

[21] M. Marcon, M. Dischinger, K. Gummadi, and A. Vahdat, "The local and global effects of traffic shaping in the internet," in *Proc. of IEEE COMSNETS*, Bangalore, India, Jan 2011.
[22] N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, "Inter-datacenter bulk transfers with netstitcher," *SIGCOMM Comput. Commun. Rev.*, vol. 41, pp. 74–85, Oct 2011.
[23] D. Niu, C. Feng, and B. Li, "Pricing cloud bandwidth reservations under demand uncertainty," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, pp. 151–162, Jun 2012.
[24] Y. He, S. Elnikety, J. Larus, and C. Yan, "Zeta: Scheduling interactive services with partial execution," in *Proc. of ACM SoCC*, San Jose, CA, Oct 2012.
[25] "Probability mass function", https://en.wikipedia.org/wiki/Probability_mass_function.
[26] B. M. Jedynak and S. Khudanpur, "Maximum likelihood set for estimating a probability mass function," *Neural Computation*, vol. 17, pp. 1508–1530, Jul 2005.
[27] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[28] E. L. Lawler and D. E. Wood, "Branch-and-bound methods: A survey," *Operations research*, vol. 14, pp. 699–719, Aug 1966.
[29] "CPLEX Optimizer", http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/.
[30] C. Joe-Wong and S. Sen, "Mathematical frameworks for pricing in the cloud: net utility, fairness, and resource allocations," *CoRR*, vol. abs/1212.0022, pp. 1–14, Jan 2012.
[31] W. Nicholson and C. Snyder, *Microeconomic theory: basic principles and extensions*. Cengage Learning, 2011.
[32] "Web hits data during WorldCup 98", http://www.acm.org/sigcomm/ITA/.