electronics

MDPI

Article

Proactive Content Delivery with Service-Tier Awareness and User Demand Prediction

Jing Hu [1], Yaling Lai [1], Ao Peng [1,*], Xuemin Hong [2] and Jianghong Shi [2]

[1] School of Information Science and Engineering, Xiamen University, Xiamen 361005, Fujian, China; 23320161153403@stu.xmu.edu.cn (J.H.); 23320151154036@stu.xmu.edu.cn (Y.L.)

[2] Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education of China, Xiamen University, Xiamen 361005, Fujian, China; xuemin.hong@xmu.edu.cn (X.H.); shijh@xmu.edu.cn (J.S.)

* Correspondence: pa@xmu.edu.cn; Tel.: +86-592-258-0150

Received: 14 November 2018; Accepted: 24 December 2018; Published: 2 January 2019

Abstract: Cost-effective delivery of massive data content is a pressing challenge facing modern mobile communication networks. In the literature, two primary approaches to tackle this challenge are service-tier differentiation and personalized proactive content caching. However, these two approaches have not been integrated and studied in a unified framework. This paper proposes an integrated proactive content delivery scheme that jointly exploits the availability of multiple service tiers and multi-user behavior prediction. Three optimal algorithms and one heuristic algorithm are introduced to solve the cost-minimization problems of multi-user proactive content delivery under different modelling assumptions. The performance of the proposed scheme is systematically investigated to reveal the impacts of proactive window size, service-tier price ratio, and traffic cost model on the system performance.

Keywords: proactive content delivery; differentiated services; redundant capacity; secondary traffic

1. Introduction

The rapid proliferation of smart phones and mobile Internet has driven an explosive growth of mobile data traffic demand. According to Cisco's report [1], global mobile data traffic will reach 49 exabytes per month by 2021. Among various types of mobile applications, content delivery (e.g., web browsing, video streaming) consumes the majority of the mobile data traffic. A Cisco report [1] estimated that video content will account for 78% of the world's total mobile traffic in 2021. However, the high price of mobile data plan (e.g., cost per Mbyte) is still one of the main factors prohibiting the ubiquitous adoption of mobile video applications. Therefore, significant research interests have been attracted in designing a mobile content delivery network that is cost-friendly to massive content delivery services.

Contradicting the high price of mobile data plan, the overall utilization of the mobile communication network's capacity is relatively low. This is because the mobile traffic demand varies significantly across space and time [2–5], while the network is typically built to accommodate the peak traffic demand. Consequently, a large amount of "redundant capacity" (i.e., the difference between the actual traffic load and the network capacity) is not used during off-peak hours [6], resulting in a low overall utilization of the network. It is widely anticipated that improving the network utilization can help to reduce the cost per bit for mobile operators and ultimately the price per bit for mobile users.

A wide range of different approaches have been studied to improve the network capacity utilization. These approaches can be broadly categorized into two types: network-centric approach and price-centric approach. The former focuses on improving the technical efficiency of the network, which can ultimately reduce the operational expenses (OPEX) and/or capital expenses (CAPEX) of mobile operators. Within this category, "green radio" [7–9] aims to dynamically adjust the number of powered-on base stations (BSs) to match the actual traffic demand, so that the OPEX (mainly the cost of electricity consumption) can be reduced. Another approach called "proactive mobile edge caching" [10–18] aims to push popular (i.e., frequently requested) content in advance and cache them in the mobile edge network or even in end-user devices, so that on-demand traffic is off-loaded to the edge network or to off-peak hours. In practice, Netflix's content delivery network (CDN), named Open Connect, can deploy servers at Internet exchange points (IXPs) and inside Internet service providers (ISPs) without operating either a backbone network or data centers, and pre-load contents on its servers during off-peak times to reduce the amount of transit traffic [19]. Furthermore, the hybrid CDN–P2P solutions, integrating P2P into the current CDN architectures, were proposed to maximize throughput and reduce expenses [20]. This load-balancing approach helps to ease the pressure of network capacity expansion, so that the CAPEX can be reduced.

The second category is the price-centric approach. The rationale is to introduce diverse communication service tiers [21,22] with differentiated prices to end users. The service tiers and prices are allowed to be changed flexibly, such that the network utilization can be improved via market dynamics. Within this category, time-based data pricing schemes [23–30] (i.e., different traffic pricing during different hours in a day) are designed to attract users through special discounts in off-peak hours. This coarse-grain approach can help to smooth the temporal variation of traffic load, but is unable to balance the spatial traffic variation. Moreover, traffic from cheaper data plans may affect the quality-of-service (QoS) of traffic from normal data plans, resulting in a degraded quality-of-experience (QoE) for normal users in off-peak hours.

An alternative price-centric approach is service-based data pricing schemes [31–38], which allow the mobile operator (i.e., mobile ISP) to offer differentiated communication service tiers associated with different prices. Paper [31] derived the optimal service qualities and associated prices for an ISP with the consideration of capacity constraints and user characteristics. Paper [32] addressed the problem of ISP service tier design based on specific requirements of the applications such as web browsing and video streaming. Financial portfolio theory was applied to develop an optimization model in [33]. Various technical, economical, and social aspects of Internet service differentiation were discussed in [34–38]. Generally speaking, compared with time-based data pricing schemes, service-based data pricing schemes offer more flexibility and greater commercial incentive. Therefore, the 5th generation (5G) mobile communication network has incorporated new technologies, such as network slicing, to enable mobile ISPs to offer differentiated service tiers.

The above-mentioned studies on mobile content delivery have mostly taken a perspective from the mobile ISPs, who are essentially data pipes and have limited knowledge about user behavior and preference. In a parallel research field of content recommendation [39–42], it has been established that the content providers (CPs), such as YouTube and Netflix, can play an active role in content delivery. The reason is that CPs hold the data of users' content preferences and historical access behavior. For general human behavior [39,40], especially for wireless data users [41,42], there is substantial evidence showing that their content consumption behaviors are fairly predictable at a fine-grain timescale (from minutes to hours). Such personalized, fine-grain information enables CPs to predict users' content demand, so that traditional proactive caching schemes can be personalized and become more effective. As a result, CP-centric personalized content delivery, as an alternative to the traditional ISP-centric content delivery, has attracted increasing research interests lately.

Previous studies on mobile content delivery have either taken an ISP-centric perspective or a CP-centric perspective. To our best knowledge, studies that unify both perspectives are still rare. In this paper, we propose a content delivery scheme that integrates both perspectives. Our scheme can simultaneously exploit the availability of differentiated services tiers and the predictability of user behavior. The main contributions of our paper are as follows. First, we propose a proactive content delivery scheme with service-tier awareness and user behavior prediction for the purpose of cost reduction. Second, considering a baseline scheme of proactive content delivery with one time-slot, we derive the optimal content delivery policy that can minimize the long-term cost. Third, considering a generalized scheme of multi-time-slot proactive content delivery, we propose a near-optimal heuristic algorithm for cost reduction. The performances of the proposed schemes are systematically evaluated to reveal key insights into the impacts of various system parameters on the cost.

The remainder of this paper is organized as follows. Section 2 describes the system model. Sections 3 and 4 formulate and analyze the problems of proactive content delivery in single-time-slot and multi-time-slot cases, respectively. Numerical results are presented in Section 5. Finally, conclusions are drawn in Section 6.

2. System Model

2.1. Model of Communication Service Tiers

We consider a system consisting of a CP, an ISP, and N users. The content data is delivered from the CP to users via the ISP, as shown in Figure 1. For simplicity, we assume that the ISP offers two service tiers: a primary traffic (PT) service and a secondary traffic (ST) service. For concreteness, we further assume that the ST only utilizes the redundant capacity of the network [6]. This assumption has two implications. First, ST has a strictly lower priority than PT, therefore the unit cost of ST (e.g., dollar per kilo bytes) is also cheaper than PT. The ratio of ST cost over PT cost is denoted as β, where $0 \leq \beta \leq 1$. Second, the capacity of ST is upper bounded by the redundant capacity of the network. The total system capacity is dependent on the infrastructure deployment and network planning of the ISP. Once a network is rolled out, the system capacity is relatively stable. Redundant capacity is given by the difference between the system capacity and the primary traffic volume. Because the primary traffic volume fluctuates over time, the redundant capacity also changes dynamically. In practice, redundant capacity can be estimated by subtracting the pre-defined system capacity by the primary traffic load, which can be measured in real-time. We note that our paper focuses on the problem of proactive content delivery, which has a time-scale of seconds or minutes. Within such a time scale, the volume of redundant capacity can be treated as fixed. Therefore, our model captures the daily traffic fluctuation by a single parameter Cr_t, which indicates the currently available redundant capacity, i.e., the upper limit for ST at time t.

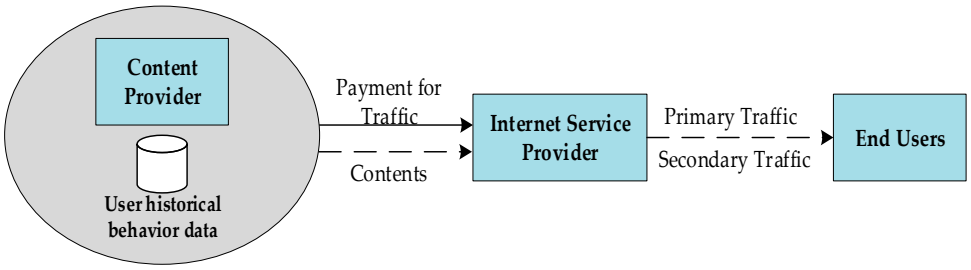


Figure 1. Illustration of the system model.

Within each service tier, the total traffic cost $C(L)$ is a function of the traffic load L. The cost is interpreted as the cost to the ISP for secondary service provision (i.e., transmit more data using redundant capacity). It is assumed that such a cost of the ISP is proportional to the cost of CP to access communication services provided by the ISP. Two cost models are considered in our paper. One is the

simple case of volume-based or linear cost, which means the cost per unit traffic remains unchanged regardless of the traffic load L. In this case, we have $C_l(L) = k_l L$, where the cost is linearly proportional to the traffic load. Another case is quadratic cost, where $C_q(L) = k_q L^2$. This is a commonly used model in the literature [18] to reflect the fact that the cost to the ISP to support higher data rates scales non-linearly with the data rate. Such a nonlinear scaling is rooted in Shannon's capacity formula: once the physical bandwidth is fixed, the data rate can be improved by increasing the transmit power, but with diminishing returns. In the literature, the cost–traffic volume function is commonly approximated by a quadratic function for analytical convenience [18].

2.2. Model of User Behavior

We assume that time is slotted into unit intervals and indexed by t. It is assumed that the CP is able to make probabilistic predictions on the users' content request behavior based on historical trace. The prediction tells that user n ($n \in \{1, 2, \ldots, N\}$) will consume a total of $\xi_{n,t}$ amount of data at time slot t with probability $p_{n,t}$, where $0 \leq \xi_{n,t} < \infty$ and $0 \leq p_{n,t} \leq 1$. A random binary variable is used to indicate whether the nth user's request actually occurs at time t, i.e.,

$$\mathrm{I}_{n,t} = \begin{cases} 1, & p_{n,t}, \\ 0, & 1 - p_{n,t} \end{cases} \tag{1}$$

It is assumed that multiple users' arrival and content consumption behaviors are independent from each other. Furthermore, user demands are assumed to be cyclic-stationary. This assumption is supported by various measurements showing that the user demand fluctuates in a periodic pattern [40,43] (e.g., on a daily basis). As a result, we can group multiple time slots into a cyclic period. The number of time slots in a period is denoted as T. It follows that

$$\xi_{n,t} = \xi_{n,t+T},\ p_{n,t} = p_{n,t+T},\ \forall n, t \tag{2}$$

2.3. Protocols of Proactive Content Delivery

We propose a protocol that is simultaneously aware of the service tiers and user behavior predictions. This requires certain degrees of collaboration and information sharing between the ISPs and CPs. At time slot t, the protocol uses the PT service tier to satisfy users' instantaneously content demand in the current slot. This is called reactive content delivery (RCD). Meanwhile, if redundant capacity is available, the protocol will proactively push a portion of the forecasted content demand in the upcoming several time slots using the ST service tier. This is called proactive content delivery (PCD). As the process iterates, the content demand at time t will be partly delivered by RCD via the PT service tier and partly by PCD via the ST service tier. Unlike traditional proactive caching schemes, the main difference here is that RCD and PCD are associated with the PT service tier and ST service tier, respectively.

Suppose that PCD is conducted over a length of W time-slots, where $1 \leq W \leq T$ and $\tau \in \{1, 2, \ldots, W\}$. When $W = 0$, the content delivery mechanism is purely reactive, which serves as our baseline case. The case of $W = 1$ is called single-slot PCD (SPCD), while the more general case of $1 < W \leq T$ is called multi-slot PCD (MPCD). We use $x_{n,t}(\tau)$ to denote the portion of data expected for user n at time $t + \tau$ but is proactively pushed to the user at time-slot t. Here τ denotes how many time slots are ahead for proactive caching. The main parameters in this paper are summarized in Table 1.

Table 1. Main parameters used in our model.

Variable	Definition
N	Number of users
T	Number of time-slots in a cyclic period
W	Window size for proactive content caching
$\xi_{n,t}$	User n's demand at time-slot t (unit: MB)
$p_{n,t}$	User n's arrival probability at time-slot t
$\mathrm{I}_{n,t}$	Random variable of user n's demand at time-slot t
Cr_t	System's redundant capacity at time-slot t (unit: MB)
$x_{n,t+1}$	Portion of proactively delivered data to be consumed at time-slot $t+1$ (unit: MB)
$x_{n,t}(\tau)$	Portion of proactively delivered data to be consumed at time-slot $t+\tau$ (unit: MB)
β	Ratio of the cost of the ST service tier over the PT tier

3. Proactive Content Delivery with Single Time-Slot

3.1. Problem Formulation

This section considers the case of proactive content delivery with single time-slot, where forecasted user demands can be sent one time-slot ahead using the ST service tier. At a given time- slot t, the cost is composed of two parts. One is the cost generated by RCD through the PT service tier, and the other part is the cost generated by PCD through the ST service tier. The time-average expected cost can be written as:

$$\eta_s(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} E\left[C\left(\sum_{n=1}^{N}(\xi_{n,t} - x_{n,t})\mathrm{I}_{n,t}\right) + \beta \cdot C\left(\sum_{n=1}^{N} x_{n,t+1}\right)\right] \tag{3}$$

where we define a $N \times T$ matrix $\mathbf{x}$, the elements of which are $x_{n,t}$, $\forall n, t$. In Equation (3), $x_{n,t+1}$ represents the portion of proactively pushed data for the next time slot t + 1, and the expectation is taken over the random variable $\mathrm{I}_{n,t}$. The received data for each user should not exceed the user's demand at time t, i.e.,

$$0 \le x_{n,t} \le \xi_{n,t} \tag{4}$$

and the total amount of proactively pushed data cannot exceed the upper limit of the redundancy capacity at the current time-slot t, i.e.,

$$\sum_{n=1}^{N} x_{n,t+1} \le Cr_t \tag{5}$$

The main objective is to minimize the total cost over the feasible space of $\mathbf{x}$. The optimization problem can be formulated as

$$\min_{\mathbf{x}} \eta_s(\mathbf{x}) \quad s.t. \begin{cases} 0 \le x_{n,t} \le \xi_{n,t} & \forall n, t \\ x_{n,t} = x_{n,t+T} & \forall n, t \\ \sum\limits_{n=1}^{N} x_{n,t} \le Cr_t & \forall n, t \\ \mathrm{I}_{n,t} \in \{0, 1\} & \forall n, t \end{cases} \tag{6}$$

For comparison purposes, also consider the baseline case of pure RCD. The time-average expected cost in this case is given by

$$\eta = \frac{1}{T}\sum_{t=1}^{T} E\left[C\left(\sum_{n=1}^{N} \xi_{n,t}\mathrm{I}_{n,t}\right)\right] \tag{7}$$

where $\sum_{n=1}^{N} \xi_{n,t}\mathrm{I}_{n,t}$ is the actual traffic load requested at time t. In this case, the system is purely reactive to the users' request and there is no decision variable to be optimized.

3.2. Linear Cost Model

Assuming the linear cost model, we can substitute $C_l(L)$ into Equation (3) to yield

$$\begin{aligned}\eta_s^l(\mathbf{x}) &= \frac{1}{T}\sum_{t=1}^{T} E\left[k_l\left(\sum_{n=1}^{N}(\xi_{n,t}-x_{n,t})\mathrm{I}_{n,t}\right)+\beta\cdot k_l\left(\sum_{n=1}^{N}x_{n,t+1}\right)\right] \\ &\overset{(a)}{=} \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N} k_l((\beta-p_{n,t})x_{n,t}+\xi_{n,t}p_{n,t})\end{aligned} \tag{8}$$

We note that the property of cyclic-stationary user demand (i.e., $x_{n,t} = x_{n,t+T}$) is used in Equation (8) to give $\sum_{t=1}^{T} x_{n,t+1} = \sum_{t=1}^{T} x_{n,t}$. From Equation (8), we can see that the optimization problem in Equation (6) becomes a linear programming problem, such that the problem can be easy solved by classic methods such as the dual interior point method.

A closer look at Equation (8) reveals a key insight that both the cost and the PCD decision variable $\mathbf{x}$ are determined by the relative difference between the cost ratio β and users' arrival probabilities $p_{n,t}$. When $p_{n,t} > \beta$, PCD for the nth user is beneficial for cost reduction; when $p_{n,t} < \beta$, PCD for the nth user becomes harmful because there is a higher likelihood that the pushed data will not be actually consumed by the user, so that the resource used for PCD is wasted. When $p_{n,t} = \beta$, PCD for the nth user makes no difference.

3.3. Quadratic Cost Model

When the cost is a quadratic function of the traffic load, the costs increase rapidly as the load increases. In this case, PCD becomes more useful because it helps to smooth the traffic load and reduce fluctuations over time. Substituting $C_q(L)$ into (3) yields:

$$\begin{aligned}\eta_s^q(\mathbf{x}) &= \frac{1}{T}\sum_{t=1}^{T} E\left[k_q\left(\sum_{n=1}^{N}(\xi_{n,t}-x_{n,t})\mathrm{I}_{n,t}\right)^2+\beta\cdot k_q\left(\sum_{n=1}^{N}x_{n,t+1}\right)^2\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} k_q\Bigg(\sum_{n=1}^{N}(\xi_{n,t}-x_{n,t})^2 p_{n,t}+\sum_{n=1}^{N}\sum_{m\neq n}(\xi_{n,t}-x_{n,t})p_{n,t}(\xi_{m,t}-x_{m,t})p_{m,t} \\ &\quad +\beta\sum_{n=1}^{N}x_{n,t+1}^2+\beta\sum_{n=1}^{N}\sum_{m\neq n}x_{n,t+1}x_{m,t+1}\Bigg) \\ &= \frac{1}{T}\sum_{t=1}^{T} k_q\Bigg(\sum_{n=1}^{N}(p_{n,t}+\beta)x_{n,t}^2+\sum_{n=1}^{N}\sum_{m\neq n}(p_{n,t}p_{m,t}+\beta)x_{n,t}x_{m,t}-2\sum_{n=1}^{N}\xi_{n,t}p_{n,t}x_{n,t} \\ &\quad -2\sum_{n=1}^{N}\sum_{m\neq n}\xi_{m,t}p_{m,t}p_{n,t}x_{n,t}+\sum_{n=1}^{N}\xi_{n,t}^2 p_{n,t}+\sum_{n=1}^{N}\sum_{m\neq n}\xi_{n,t}\xi_{m,t}p_{n,t}p_{m,t}\Bigg)\end{aligned} \tag{9}$$

We can see that in this case, we no longer have a simple intuitive solution for $\mathbf{x}$. However, it can be proved that the problem in Equation (9) is a convex optimization problem (see Appendix A). Hence, the optimal solution can be readily solved using standard convex optimization techniques.

4. Proactive Content Delivery with Multiple Time-Slots

4.1. Problem Formulation

As a generalization from the single-time slot case, portions of user's predicted demand can be pushed to users by multiple time-slots ahead through the ST service tier. The time-average expected cost in this case is given by:

$$\eta_m(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} E\left[C\left(\sum_{n=1}^{N}\left(\xi_{n,t} - \sum_{\tau=1}^{W} x_{n,t-\tau}(\tau)\right)\mathrm{I}_{n,t}\right) + \beta \cdot C\left(\sum_{n=1}^{N}\sum_{\tau=1}^{W} x_{n,t}(\tau)\right)\right] \tag{10}$$

where the decision variable $\mathbf{x}$ is a $N \times T \times W$ matrix, whose elements are given by $x_{n,t}(\tau),\ \forall n, t, \tau$. Here, what differs from the single-time-slot case is that user n's cached data at time t is the accumulated data pushed from the previous W time-slots. PCD for each user is constrained by the individual user demand, i.e.,

$$\begin{aligned} x_{n,t-\tau}(\tau) &\geq 0, \\ \sum_{\tau=1}^{W} x_{n,t-\tau}(\tau) &\leq \xi_{n,t} \end{aligned} \tag{11}$$

in addition, the total amount of PCD data of all users at any time-slot t cannot exceed the current redundant capacity, i.e.,

$$\sum_{n=1}^{N}\sum_{\tau=1}^{W} x_{n,t}(\tau) \leq Cr_t \tag{12}$$

the optimization problem can then be formulated as:

$$\min_{\mathbf{x}} \eta_m(\mathbf{x})$$
$$s.t. \begin{cases} x_{n,t}(\tau) \geq 0 & \forall n, t, \tau \\ x_{n,t}(\tau) = x_{n,t+T}(\tau) & \forall n, t, \tau \\ \sum_{\tau=1}^{W} x_{n,t-\tau}(\tau) \leq \xi_{n,t} & \forall n, t, \tau \\ \sum_{n=1}^{N}\sum_{\tau=1}^{W} x_{n,t}(\tau) \leq Cr_t & \forall n, t, \tau \\ \mathrm{I}_{n,t} \in \{0, 1\} & \forall n, t \end{cases} \tag{13}$$

4.2. Linear Cost Model

Substituting the linear cost function $C_l(L)$ into Equation (10) we get

$$\begin{aligned} \eta_m^l(\mathbf{x}) &= \tfrac{1}{T}\sum_{t=1}^{T} E\left[C_l\left(\sum_{n=1}^{N}\left(\xi_{n,t} - \sum_{\tau=1}^{W} x_{n,t-\tau}(\tau)\right)\mathrm{I}_{n,t}\right) + \beta \cdot C_l\left(\sum_{n=1}^{N}\sum_{\tau=1}^{W} x_{n,t}(\tau)\right)\right] \\ &= \tfrac{1}{T}\sum_{t=1}^{T} k_l\left(\sum_{n=1}^{N}\left(\xi_{n,t} - \sum_{\tau=1}^{W} x_{n,t-\tau}(\tau)\right)p_{n,t}\right) + \beta \cdot k_l\left(\sum_{n=1}^{N}\sum_{\tau=1}^{W} x_{n,t}(\tau)\right) \\ &\overset{(b)}{=} \tfrac{1}{T}\sum_{t=1}^{T} k_l \sum_{n=1}^{N}\left(\sum_{\tau=1}^{W} x_{n,t}(\tau)(\beta - p_{n,t}) + \xi_{n,t}p_{n,t}\right) \end{aligned} \tag{14}$$

In (14), the equality (b) follows by $\sum_{t=1}^{T}\sum_{\tau=1}^{W} x_{n,t-\tau}(\tau) = \sum_{t=1}^{T}\sum_{\tau=1}^{W} x_{n,t}(\tau)$. We can see that the optimization problem reduces to a linear programing problem. Similar to the case of single time-slot, the effectiveness of PCD still depends on the relative difference between the traffic cost ratio β and user n's arrival probability $p_{n,t}$. However, the proactive data user n received from different time-slot, i.e., $x_{n,t-\tau}(\tau)$, depends on the redundant capacity of the previous W time-slots. This requires proper monitoring of real-time redundant capacity over multiple time slots.

4.3. Quadratic Cost Model

Substituting the quadratic cost function $C_q(L)$ into Equation (10), we have

$$\begin{aligned}
\eta_m^q(\mathbf{x}) &= \tfrac{1}{T}\sum_{t=1}^{T} E\left[k_q\left(\sum_{n=1}^{N}\left(\xi_{n,t}-\sum_{\tau=1}^{W}x_{n,t-\tau}(\tau)\right)\mathrm{I}_{n,t}\right)^2+\beta k_q\left(\sum_{n=1}^{N}\sum_{\tau}^{W}x_{n,t}(\tau)\right)^2\right]\\
&= \tfrac{1}{T}\sum_{t=1}^{T}\Bigg\{E\Bigg[k_q\sum_{n=1}^{N}\left(\xi_{n,t}-\sum_{\tau=1}^{W}x_{n,t-\tau}(\tau)\right)^2\mathrm{I}_{n,t}^2\\
&\quad +k_q\sum_{n=1}^{N}\sum_{n\neq m}\left(\xi_{n,t}-\sum_{\tau=1}^{W}x_{n,t-\tau}(\tau)\right)\left(\xi_{m,t}-\sum_{\tau=1}^{W}x_{m,t-\tau}(\tau)\right)\mathrm{I}_{n,t}\mathrm{I}_{m,t}\Bigg]+\beta k_q\left(\sum_{n=1}^{N}\sum_{\tau}^{W}x_{n,t}(\tau)\right)^2\Bigg\} \qquad (15)\\
&= \tfrac{1}{T}\sum_{t=1}^{T}\Bigg\{k_q\sum_{n=1}^{N}\left(\xi_{n,t}-\sum_{\tau=1}^{W}x_{n,t-\tau}(\tau)\right)^2 p_{n,t}\\
&\quad +k_q\sum_{n=1}^{N}\sum_{n\neq m}\left(\xi_{n,t}-\sum_{\tau=1}^{W}x_{n,t-\tau}(\tau)\right)\left(\xi_{m,t}-\sum_{\tau=1}^{W}x_{m,t-\tau}(\tau)\right)p_{n,t}p_{m,t}+\beta k_q\left(\sum_{n=1}^{N}\sum_{\tau}^{W}x_{n,t}(\tau)\right)^2\Bigg\}
\end{aligned}$$

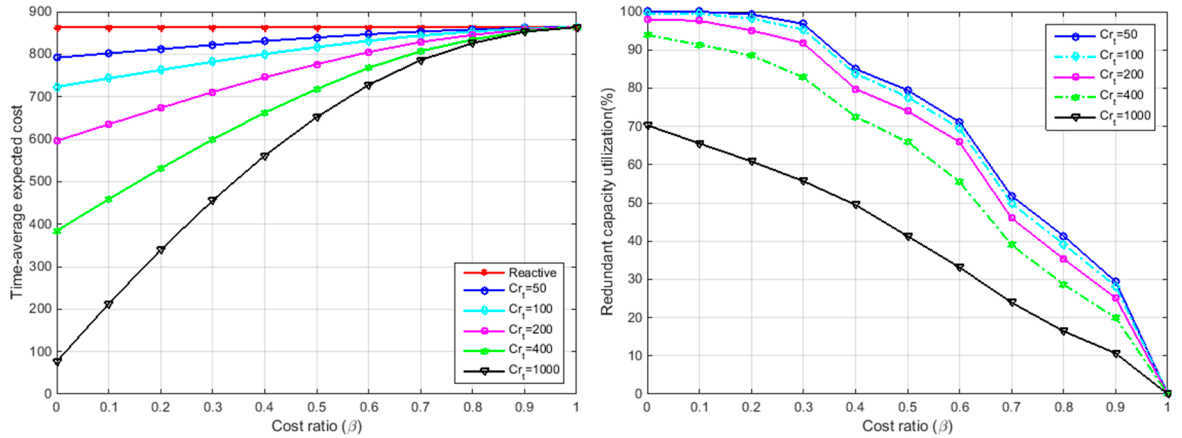
This yields a complicated non-linear optimization problem and there is no straightforward proof for its convexity. However, because the utility function can be easily evaluated in closed-form, general purpose heuristic search algorithms such as the pattern search [44] can be used to solve the problem effectively.

5. Simulation Results

This section presents numerical results to our previous analysis. For illustration purposes, we set $T = 10$ and $N = 3$. User n's demand at time t is drawn from a uniform distribution on [0, 500]; the arrival probability of user n at time t follows a uniform distribution on [0, 1]. The scaling constants in the linear and quadratic cost models are given by $k_l = 2$ and $k_q = 0.005$, respectively. The case of pure RCD, where there is no proactive caching, is also presented as a performance benchmark.

5.1. Case of Single Time-Slot

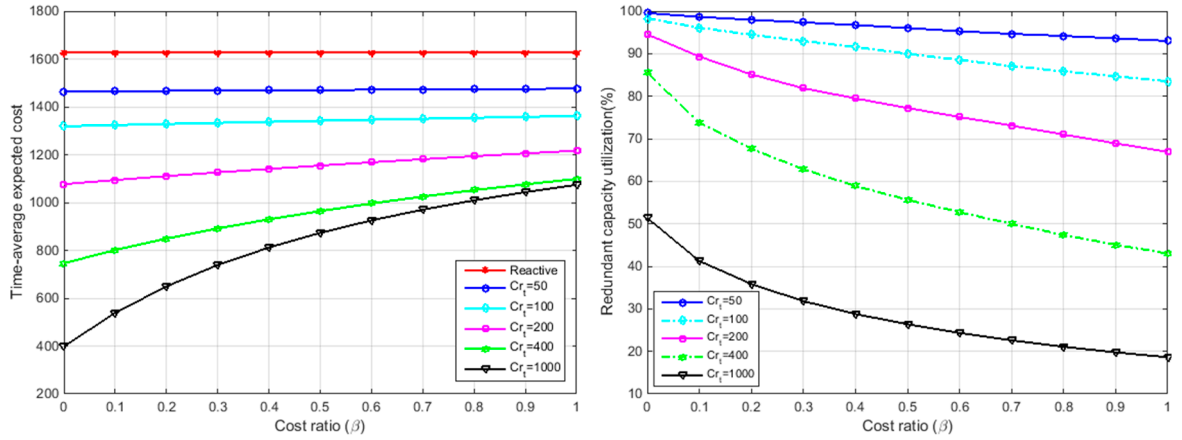
Using the linear cost model, Figure 2 shows how the time-average expected cost and the redundant capacity utilization changes as a function of the ST/PT cost ratio β. The results are obtained by solving the linear optimization problem defined in Section 3.2 and averaging over 100 realizations. It is observed that a smaller value of β leads to a lower cost and a higher utilization of the redundant capacity. This is expectable because a smaller value of β would better encourage the use of PCD using the ST service tier. When $\beta = 1$, which means the two service tiers have the same cost, there is no performance gain to use PCD at all. Moreover, we can see that larger amount of redundant capacity helps to reduce the cost because more user demand can be accommodated via the ST service tier.



(**a**) (**b**)

Figure 2. (**a**) The time-average expected cost as a function of the ST/PT cost ratio β; (**b**) the redundant capacity utilization as a function of the ST/PT cost ratio β (linear cost model, varying redundant capacity Cr_t).

Using the quadratic cost model, Figure 3 shows how the time-average expected cost and the redundant capacity utilization changes as a function of the ST/PT cost ratio β. The results are obtained by solving the convex optimization problem defined in Section 3.3. The general trend observed in Figure 3 is similar to that in Figure 2, i.e., a smaller value of β leads to a lower cost and a higher utilization of the redundant capacity. However, a key difference to Figure 2 occurs when β approaches 1, where PCD is shown to be useful for cost reduction even when the cost of ST and PT are the same. For example, at Cr_t = 400 and β = 1, the time-average cost can be reduced by nearly 32% (as opposed to 0% in Figure 2) and the redundant traffic utilization is about 43% (as opposed to 0% in Figure 2). This is because the PCD can help to smooth the user demand in time, while a more balanced user demand yields a lower cost under the quadratic cost model.



(**a**) (**b**)

Figure 3. (**a**) The time-average expected cost as a function of the ST/PT cost ratio β; (**b**) the redundant capacity utilization as a function of the ST/PT cost ratio β (quadratic cost model, varying redundant capacity Cr_t).

5.2. Case of Multiple Time-Slot

Figure 4 shows three figures related to the performance of multi-time-slot PCD under the linear cost model. The results are obtained by solving the linear optimization problem defined in Section 4.2.

The general conclusions drawn from Figure 4 are the same as that in Figure 2, i.e., PCD is not useful when there is no cost difference between ST and PT service tiers. Apart from this, Figure 4a–c further reveal the impact of proactive window size on the performance. It is observed that increasing the window size does help to further reduce the cost, but the improvement is limited and becomes insignificant when W is greater than five. In Figure 4c, we can see that when the value of β increases, the effectiveness of cost reduction by increasing W decreases. This suggests that when the costs of the two service tiers are comparable, increasing the proactive window size W will become less effective for cost saving.

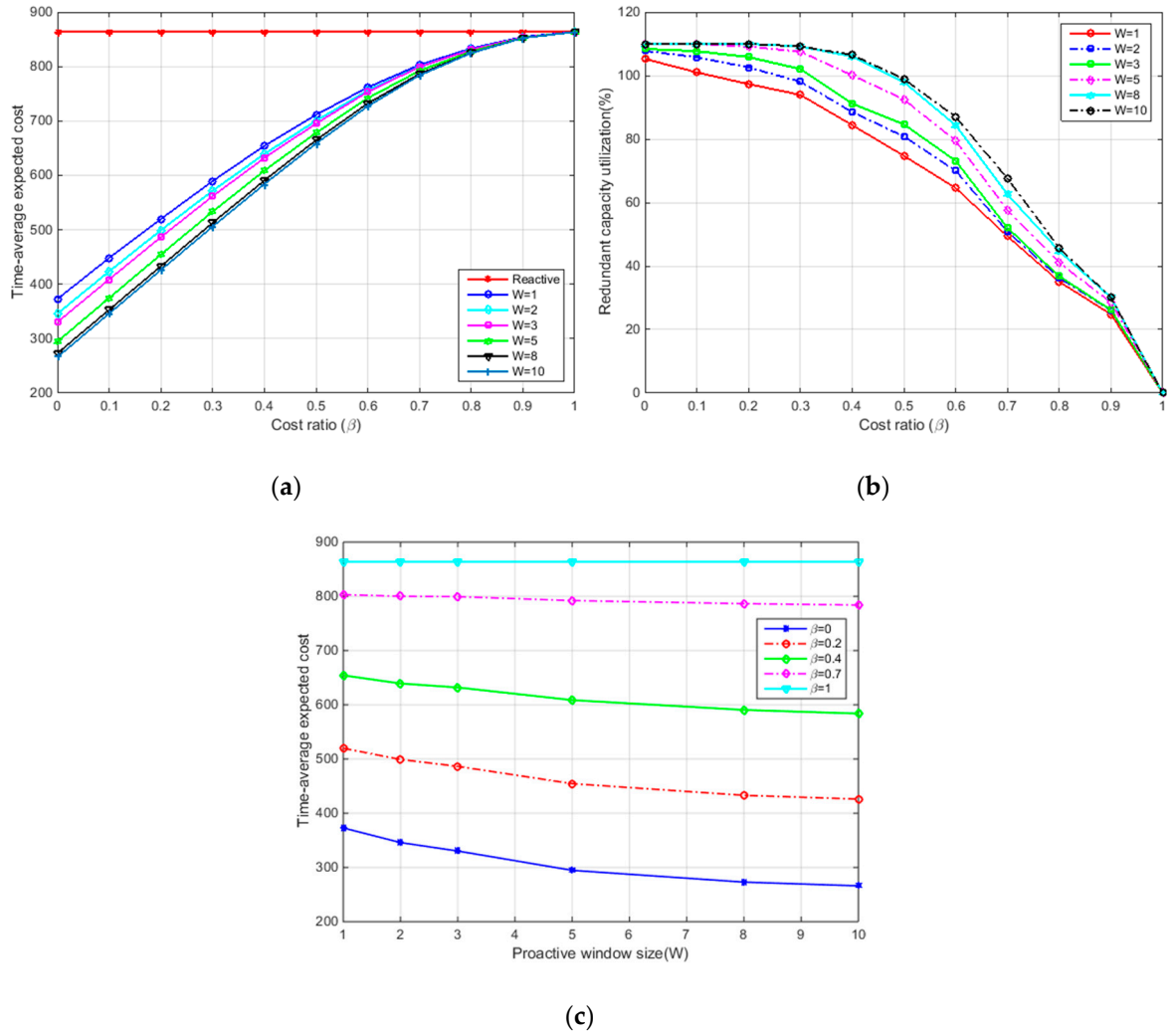


Figure 4. (**a**) The time-average expected cost as a function of the ST/PT cost ratio β; (**b**) the redundant capacity utilization as a function of the ST/PT cost ratio β; (**c**) the time-average expected cost as a function of the proactive window size W (linear cost model, $Cr_t = 400$, varying window size W).

Finally, Figure 5 shows three figures related to the performance of multi-time-slot PCD under the quadratic cost model. The results are obtained by solving the non-linear optimization problem defined in Section 4.3 using pattern search. Compared with Figure 4, Figure 5 shows that using PCD is always useful for cost reduction regardless of the values of β. Even when $\beta = 1$, the cost can still be reduced by 53% thanks to the load smoothing effect. Moreover, increasing the window size also helps for load smoothing, and is hence considered beneficial for all values of β. Table 2 further demonstrates the smoothing effect of multi-time-slot PCD on network traffic load. Given $Cr_t = 400$ and $\beta = 0.5$, the variances of the actual traffic across different time slots is shown as a function of the window size. We

can see that increasing W helps to reduce the variance of the traffic load, but has diminishing returns especially when W becomes greater than five.

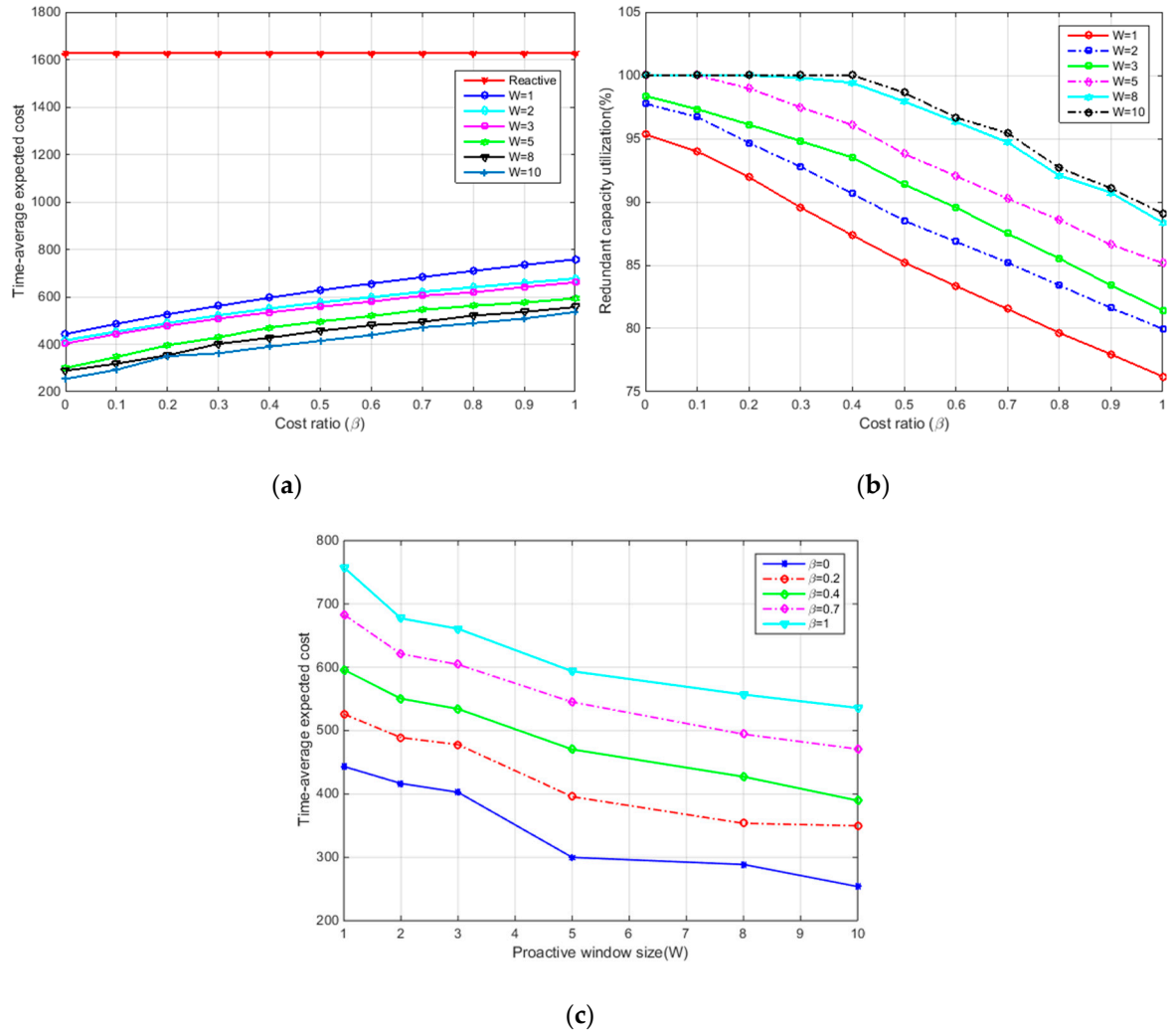


Figure 5. (**a**) The time-average expected cost as a function of the ST/PT cost ratio β; (**b**) the redundant capacity utilization as a function of the ST/PT cost ratio β; (**c**) the time-average expected cost as a function of the proactive window size W (quadratic cost model, $Cr_t = 400$, varying window size W).

Table 2. The variance of traffic demand.

W	1	2	3	5	8	10
Variance ($\times 10^4$)	8.3188	5.7479	5.0893	4.2842	3.4933	3.3491

The above simulation results show that both single time-slot and multiple time-slot PCD can bring good performance gain for CP. The performance gain increases with lower cost rate β and larger window size W. However, the performance gain is fundamentally constrained by the volume of redundant capacity. In practice, this means close cooperation must be established between CP and ISP so that the volume of redundant capacity in the current network can be measured and shared in real time. For the ISP, our model helps to improve the overall utilization of network infrastructure and generate additional revenue. For CP, our model helps to attract users and promote content consumption by reducing the cost of content delivery per bit. In summary, our model can offer a win-win situation for ISP and CP.

6. Conclusions

This paper proposes a personalized PCD scheme that aims to minimize the total cost of content delivery by means of multiple service-tier transmission and multi-user behavior prediction. The problem of personalized PCD has been systematically investigated in single-time-slot and multi-time-slot cases, under both linear and quadratic cost models. Three optimal algorithms and one heuristic algorithm have been presented to solve the respective optimization problems. Simulation results have demonstrated the effectiveness of the proposed PCD scheme and revealed the impacts of proactive window size, service-tier cost ratio, and traffic cost model on the cost of content delivery. We conclude that personalized PCD over multiple service tiers can effectively reduce the cost when the cost is sensitive to the total traffic load and/or the type of service tiers.

Author Contributions: Conceptualization, J.H. and Y.L.; Formal analysis, J.H.; Investigation, Y.L.; Methodology, A.P.; Project administration, A.P. and X.H.; Resources, J.S.; Supervision, J.S.; Validation, Y.L.; Writing—Original draft, J.H.; Writing—Review & editing, A.P. and X.H.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) (Grant No.: 61571378) and the National Key Research and Development Program of China (Grant No.: 2018YFB0505202).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Substituting Equation (9) into Equation (6), the problem of single-slot cost optimization becomes:

$$\begin{aligned}\min_{\mathbf{x}} \frac{1}{T}\sum_{t=1}^{T}\Bigg\{&\sum_{n=1}^{N} k_q(\beta+p_{n,t})x_{n,t}^2+\sum_{n=1}^{N}\sum_{m\neq n} k_q(p_{n,t}p_{m,t}+\beta)x_{n,t}x_{m,t}\\&-\sum_{n=1}^{N} 2k_q\left(\xi_{n,t}p_{n,t}+\sum_{m\neq n}\xi_{m,t}p_{n,t}p_{m,t}\right)x_{n,t}\Bigg\}+cons\\ s.t.&\begin{cases}0\le x_{n,t}\le \xi_{n,t} & \forall n,t\\ x_{n,t}=x_{n,t+T} & \forall n,t\\ \sum_{n=1}^{N} x_{n,t}\le Cr_t & \forall n,t\\ \mathrm{I}_{n,t}\in\{0,1\} & \forall n,t\end{cases}\end{aligned} \tag{A1}$$

where, $cons=\frac{k_q}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\left(\xi_{n,t}^2p_{n,t}+\sum_{m\neq n}\xi_{n,t}\xi_{m,t}p_{n,t}p_{m,t}\right)$. The value of *cons* mainly depends on $p_{n,t}$ and $\xi_{n,t}$, so *cons* is independent of the variables and not relevant for the minimization of the objective function. We can see this is a quadratic programming problem. In order to prove the convexity of the objective function, we define $\mathbf{Q}_t$ as its Hessian matrix, whose elements are given by:

$$\begin{aligned}&[\mathbf{Q}_t]_{n,n}=k_q(p_{n,t}+\beta)\\&[\mathbf{Q}_t]_{m,n}=k_q(p_{m,t}p_{n,t}+\beta),\ n\neq m\\&[\mathbf{q}_t]_{n}=-2k_q\left(\xi_{n,t}p_{n,t}+\sum_{m\neq n}\xi_{m,t}p_{n,t}p_{m,t}\right)\end{aligned} \tag{A2}$$

Proof: suppose that $[\widetilde{\mathbf{Q}}_t]_{n,n}=p_{n,t}$, $[\widetilde{\mathbf{Q}}_t]_{m,n}=p_{m,t}p_{n,t}$ ($n\neq m$ and $n,m\in\{1,2,\ldots,N\}$), then:

$$\widetilde{\mathbf{Q}}_t=\mathbf{P}_t^T\breve{\mathbf{Q}}_t\mathbf{P}_t \tag{A3}$$

where, $\mathbf{P}_t=\mathrm{diag}\{p_{1,t},\ldots,p_{N,t}\}$, $\left[\breve{\mathbf{Q}}_t\right]_{nn}=\frac{1}{p_{n,t+1}}$, $\left[\breve{\mathbf{Q}}_t\right]_{nm}=1$. Let us set vector $\mathbf{b}=\begin{bmatrix}\sqrt{\beta} & \sqrt{\beta} & \ldots & \sqrt{\beta}\end{bmatrix}^T$, then we have $\hat{\mathbf{Q}}_t=\widetilde{\mathbf{Q}}_t+\mathbf{b}\mathbf{b}^T$, and:

$$\mathbf{Q}_t=k_q\hat{\mathbf{Q}}_t \tag{A4}$$

Therefore whether $\mathbf{Q}_t$ is a positive definite matrix can be determined by the nature of $\hat{\mathbf{Q}}_t$. First, $\widetilde{\mathbf{Q}}_t$ can be proved to be a positive definite matrix. Here g_n is defined as the n-order principal minor determinant of $\widetilde{\mathbf{Q}}_t$, $n \in \{1, 2, \ldots, N\}$, i.e.,

$$g_n = \begin{vmatrix} \alpha_1 & 1 & \cdots & 1 \\ 1 & \alpha_2 & & \\ \vdots & & \ddots & \\ 1 & & & \alpha_n \end{vmatrix} \tag{A5}$$

where $\alpha_n = \frac{1}{p_{n,t+1}} > 1$, and then:

$$g_n = \left(\alpha_1 + \sum_{k=2}^{n} \frac{1-\alpha_1}{1-\alpha_k}\right) \prod_{k=2}^{n} (\alpha_k - 1) \tag{A6}$$

We have $g_n > 0, \forall n$, i.e., all principal minor determinants of $\breve{\mathbf{Q}}_t$ are positive, thereby $\breve{\mathbf{Q}}_t$ is a positive definite matrix, meaning that $\breve{\mathbf{Q}}_t$ is also positive definite. According to Sylvester's theorem:

$$\det(\mathbf{X} + \mathbf{cr}) = \det(\mathbf{X})(1 + \mathbf{r}\mathbf{X}^{-1}\mathbf{c}) = \det(\mathbf{X}) + \mathbf{r}adj(\mathbf{X})\mathbf{c} \tag{A7}$$

we can write:

$$\left|\hat{\mathbf{Q}}_\mathbf{t}\right| = \left|\widetilde{\mathbf{Q}}_t + \mathbf{b}\mathbf{b}^T\right| = \left|\widetilde{\mathbf{Q}}_t\right|\left|1 + \mathbf{b}^T\widetilde{\mathbf{Q}}_t{}^{-1}\mathbf{b}\right| \tag{A8}$$

Because $\widetilde{\mathbf{Q}}_t$ is positive definite as shown above, we can get $\left|\hat{\mathbf{Q}}_t\right| > 0$, therefore it can be concluded that the objective function's Hessian matrix $\mathbf{Q_t}$ is a positive definite matrix, which ends our proof that the optimization problem in Equation (A1) is a convex quadratic programming problem.

References

1. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021. Available online: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html (accessed on 23 January 2018).
2. Lee, D.; Zhou, S.; Zhong, X.; Niu, Z.; Zhou, X.; Zhang, H. Spatial modeling of the traffic density in cellular networks. *IEEE Wirel. Commun.* **2014**, *21*, 80–88. [CrossRef]
3. Zhao, Z.; Li, M.; Li, R.; Zhou, Y. Temporal-spatial distribution nature of traffic and base stations in cellular networks. *IET Commun.* **2017**, *11*, 2410–2416. [CrossRef]
4. Leland, W.E.; Taqqu, M.S.; Willinger, W.; Wilson, D.V. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. Netw.* **1995**, *2*, 1–15. [CrossRef]
5. Park, K.; Willinger, W. Self-similar network traffic: An overview. In *Self-Similar Network Traffic and Performance Evaluation*; Park, K., Willinger, W., Eds.; John Wiley and Sons: New York, NY, USA, 2000; pp. 1–38, ISBN 978-0-471-31974-0.
6. Zhang, Y.; Lu, H.; Wang, H.; Hong, X. Cognitive cellular content delivery networks: Cross-layer design and analysis. In Proceedings of the IEEE Vehicular Technology Conference (VTC spring), Nanjing, China, 15–18 May 2016; pp. 1–6.
7. Wang, B.; Wu, Y.; Liu, K.J.R.; Clancy, T.C. An anti-jamming stochastic game for cognitive radio networks. *IEEE J. Sel. Areas Commun.* **2011**, *29*, 877–889. [CrossRef]
8. Han, C.; Harrold, T.; Armour, S.; Krikidis, I.; Videv, S.; Grant, P.M.; Haas, H.; Thompson, J.S.; Ku, I.; Wang, C.; et al. Green radio: Radio techniques to enable energy-efficient wireless networks. *IEEE Commun. Mag.* **2011**, *49*, 46–54. [CrossRef]
9. Ismail, M.; Zhuang, W. Green radio communications in a heterogeneous wireless medium. *IEEE Wirel. Commun.* **2014**, *21*, 128–135. [CrossRef]
10. Kangasharju, J.; Roberts, J.; Ross, K.W. Object replication strategies in content distribution networks. *Comput. Commun.* **2002**, *25*, 376–383. [CrossRef]

11. Vakali, A.; Pallis, G. Content delivery networks: Status and trends. *IEEE Internet Comput.* **2003**, *7*, 68–74. [CrossRef]
12. Pallis, G.; Vakali, A. Insight and perspectives for content delivery networks. *Commum. ACM* **2006**, *49*, 101–106. [CrossRef]
13. Ahlehagh, H.; Dey, S. Video-aware scheduling and caching in the radio access network. *IEEE/ACM Trans. Netw.* **2014**, *22*, 1444–1462. [CrossRef]
14. Ahlehagh, H.; Dey, S. Video caching in radio access network: Impact on delay and capacity. In Proceedings of the IEEE Wireless Communications and Network Conference (WCNC), Paris, France, 1–4 April 2012; pp. 2276–2281.
15. Xu, Y.; Li, Y.; Wang, Z.; Lin, T. Coordinated caching model for minimizing energy consumption in radio access network. In Proceedings of the IEEE International Conference on Communications (ICC), Sydney, Australia, 10–14 June 2014; pp. 2406–2411.
16. Shoukry, O.; Elmohsen, M.A.; Tadrous, J.; Gamal, H.E.; Elbatt, T.; Wanas, N.; Elnakieb, Y.; Khairy, M. Proactive scheduling for content pre-fetching in mobile networks. In Proceedings of the IEEE International Conference on Communications (ICC), Sydney, Australia, 10–14 June 2014; pp. 2848–2854.
17. Shoukry, O.K.; Fayek, M.B. Evolutionary scheduler for content pre-fetching in mobile networks. In Proceedings of the 2013 AAAI Fall Symposium Series, Arlington, VA, USA, 15–17 November 2013; pp. 386–391.
18. Tadrous, J.; Eryilmaz, A.; Gamal, H.E. Joint smart pricing and proactive content caching for mobile services. *IEEE/ACM Trans. Netw.* **2016**, *24*, 2357–2371. [CrossRef]
19. Bottger, T.; Cuadrado, F.; Tyson, G.; Castro, I.; Uhlig, S. Open connect everywhere: A glimpse at the internet ecosystem through the lens of the Netflix CDN. *ACM SIGCOMM Comp. Commun. Rev.* **2018**, *48*, 28–34. [CrossRef]
20. Hasslinger, G.; Hartleb, F. Content delivery and caching from a network provider's perspective. *Comput. Netw.* **2011**, *55*, 3991–4006. [CrossRef]
21. Kimbler, K.; Taylor, M. Value added mobile broadband services innovation driven transformation of the 'smart pipe'. In Proceedings of the 2012 16th International Conference on Intelligence in Next Generation Networks, Berlin, Germany, 8–11 October 2012; pp. 30–34.
22. Yang, Z.; Ma, Z. Analysis of communication operators transformation on smart pipe. In Proceedings of the 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 26–27 August 2013; pp. 131–134.
23. Jiang, L.; Parekh, S.; Walrand, J. Time-dependent network pricing and bandwidth trading. In Proceedings of the NOWS Workshops 2008-IEEE Network Operations and Management Symposium Workshops, Salvador da Bahia, Brazil, 7–11 April 2008; pp. 193–200.
24. Joe-Wong, C.; Ha, S.; Chiang, M. Time-dependent broadband pricing: Feasibility and benefits. In Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS), Minneapolis, MN, USA, 20–24 June 2011; pp. 288–298.
25. Sen, S.; Joe-Wong, C.; Ha, S.; Chiang, M. Smart data pricing: Using economics to manage network congestion. *Commun. ACM* **2015**, *58*, 86–93. [CrossRef]
26. Zhang, L. Smart Data Pricing in Wireless Data Networks: An Economic Solution to Congestion. Ph.D. Thesis, Hong Kong Polytechnic University, Hong Kong, China, 2016.
27. Zhang, L.; Wu, W.J.; Wang, D. Time dependent pricing in wireless data networks: Flat-rate vs. Usage-based schemes. In Proceedings of the 33rd IEEE Annual Conference on Computer Communications (IEEE INFOCOM), Toronto, ON, Canada, 27 April–2 May 2014; pp. 700–708.
28. Kesidis, G.; Das, A.; de Veciana, G. On flat-rate and usage-based pricing for tiered commodity internet services. In Proceedings of the 42nd Annual Conference on Information Sciences and Systems, Princeton, NJ, USA, 19–21 March 2008; pp. 304–308.
29. Chau, C.K.; Wang, Q.; Chiu, D.M. On the Viability of Paris Metro Pricing for Communication and Service Networks. In Proceedings of the Conference on IEEE INFOCOM, San Diego, CA, USA, 15–19 March 2010; pp. 1–9.
30. Ma, R.T.B. Usage-Based Pricing and Competition in Congestible Network Service Markets. *IEEE/ACM Trans. Netw.* **2016**, *24*, 3084–3097. [CrossRef]

31. Zou, M.; Ma, R.T.B.; Wang, X.; Xu, Y. On optimal service differentiation in congested network markets. In Proceedings of the IEEE Conference on Computer Communications (INFOCOM), Atlanta, GA, USA, 1–4 May 2017; pp. 1–9.
32. Dai, W.; Jordan, S. ISP Service Tier Design. *IEEE/ACM Trans. Netw.* **2016**, *24*, 1434–1447. [CrossRef]
33. Nesse, P.J.; Gaivoronski, A.; Lonsethagen, H. Ecosystem, QoE and pricing of end to end differentiated services. In Proceedings of the 6th International Conference on Information, Intelligence, Systems and Applications (IISA), Corfu, Greece, 6–8 July 2015; pp. 1–7.
34. Gibbens, R.; Mason, R.; Steinberg, R. Internet service classes under competition. *IEEE J. Sel. Areas Commun.* **2000**, *18*, 2490–2498. [CrossRef]
35. Ma, R.T.B.; Misra, V. The public option: A nonregulatory alternative to network neutrality. *IEEE/ACM Trans. Netw.* **2013**, *21*, 1866–1879. [CrossRef]
36. Wang, S.; Xuan, D.; Bettati, R.; Zhao, W. Providing absolute differentiated services for real-time applications in static-priority scheduling networks. *IEEE/ACM Trans. Netw.* **2004**, *12*, 326–339. [CrossRef]
37. Nandy, B.; Ethridge, J.; Lakas, A.; Chapman, A. Aggregate flow control: Improving assurances for differentiated services network. In Proceedings of the 20th Annual Joint Conference of the IEEE-Computer-Society/IEEE-Communication-Society, Anchorage, AK, USA, 22–26 April 2001; pp. 1340–1349.
38. Bouyoucef, K.; Khorasani, K. A robust distributed congestion-control strategy for differentiated-services network. *IEEE Trans. Ind. Electron.* **2009**, *56*, 608–617. [CrossRef]
39. Farrahi, K.; Gatica-Perez, D. Discovering human routines from cell phone data with topic models. In Proceedings of the 12th IEEE International Symposium on Wearable Computers, Pittsburgh, PA, USA, 28 September–1 October 2008; pp. 29–32.
40. Song, C.; Barabási, A.L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021. [CrossRef] [PubMed]
41. Fikir, O.B.; Yaz, I.O.; Ozyer, T. A movie Rating Prediction Algorithm with Collaborative Filtering. In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Odense, Denmark, 9–11 August 2010; pp. 321–325.
42. Salter, J.; Antonopoulos, N. CinemaScreen Recommender Agent: Combining Collaborative and Content-Based Filtering. *IEEE Intell. Syst.* **2006**, *21*, 35–41. [CrossRef]
43. Feknous, M.; Houdoin, T.; Guyader, B.L.; Biasio, J.D.; Gravey, A.; Gijon, J.A.T. Internet traffic analysis: A case study from two major European operators. In Proceedings of the 2014 IEEE Symposium on Computers and Communication (ISCC), Funchal, Portugal, 23–26 June 2014; pp. 1–7.
44. Lewis, R.M.; Torczon, V. Pattern search methods for linearly constrained minimization. *SIAM J. Optim.* **2000**, *10*, 917–941. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).