Pay as Your Service Needs: An Application-Driven Pricing Approach for the Internet Economics

HONG XIE, Chongqing University, China
WEIJIE WU, ZeroEx Inc., USA
RICHARD T. B. MA, National University of Singapore, Singapore
JOHN C. S. LUI, The Chinese University of Hong Kong, Hong Kong

Various differentiated pricing schemes have been proposed for the Internet market. Aiming at replacing the traditional single-class pricing for better welfare, yet, researchers have shown that existing schemes can bring only marginal profit gain for the ISPs. In this article, we point out that a proper form of differentiated pricing for the Internet should not only consider congestion, but more importantly, it should provide *application specific* treatment to data delivery. Formally, we propose an "application-driven pricing" approach, where an ISP offers a number of service classes in terms of a guaranteed quality of service and announces a unit usage price for each class, and content providers are free to choose which class to use depending on the requirement of their applications. Unlike previous studies, we point out that the revenue gain of multi-class pricing under our scheme can be significant. This is because we capture important aspects of application heterogeneity and take the quality of service and price as control knobs. We identify key factors that impact the revenue gain and reveal fundamental understandings on *when* and *why* an application-driven multi-class pricing can *significantly* increase the revenue of ISPs.

CCS Concepts: • **Networks** → *Network protocols*; *Network services*;

Additional Key Words and Phrases: Application driven pricing, quality of service, revenue maximization

ACM Reference format:
Hong Xie, Weijie Wu, Richard T. B. Ma, and John C. S. Lui. 2019. Pay as Your Service Needs: An Application-Driven Pricing Approach for the Internet Economics. *ACM Trans. Internet Technol.* 19, 4, Article 52 (November 2019), 28 pages.
https://doi.org/10.1145/3361148

1 INTRODUCTION

The Internet market is typically considered as a two-sided market, where Internet service providers (ISPs) stand in between end-users and content providers (CPs). They charge end-users for Internet access on the one side and CPs for content delivery on the other side. The traditional charging schemes of ISPs are mainly based on the volume of traffic [11] or the 95-percentile of bandwidth

John C.S. Lui is partially supported by the GRF-14200117.
Authors' addresses: H. Xie, College of Computer Science, Chongqing University, No.174 Shazhengjie, Shapingba, Chongqing, China; email: xiehong2018@cqu.edu.cn; W. Wu, ZeroEx Inc., 300 Beale St, San Francisco, CA, USA; email: wuwjpku@gmail.com; R. T. B. Ma, School of Computing, National University of Singapore, 15 Computing Drive, Republic of Singapore; email: tbma@comp.nus.edu.sg; J. C. S. Lui, Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin N.T., Hong Kong; email: cslui@cse.cuhk.edu.hk.

1533-5399/2019/11-ART52 $15.00
https://doi.org/10.1145/3361148

capacity [25]. They indeed achieved great economical success in the past, thanks to the tremendous growth of user population and explosion of numerous applications over the Internet. However, in recent years, ISPs are facing increasing pressure in revenue growth. Because the increase of the user population has been slowing down significantly, ISPs need to find alternative means of economic growth. More importantly, as applications are becoming more heterogeneous in nature, it becomes more problematic to transit traffic for them using a single channel without any quality guarantee. In fact, applications may differ in many aspects: the amount of traffic usage, the quality requirement, the reservation price (or referred to as "willingness to pay price"), and so on. Using a single channel to serve all applications leads to a common non-guaranteed service quality, with some applications exceeding their real requirement causing resource wastage, while others being in deficit of service quality. This motivates, or in fact forces, some giant CPs (e.g., Facebook, Google, Netflix) to build their own data center interconnections and CDN infrastructures [4, 34] to deliver their contents, taking away a lot of business that conventionally belong to ISPs.

We believe this situation is not merely a dilemma, but it implies potential business growth for ISPs. On the one hand, it is not CPs' expertise to build networks. If ISPs can do it (and potentially, with lower constructing and operating costs), why won't CPs be willing to buy such services? On the other hand, if ISPs can provide individualized services for various applications, it becomes possible to use differentiated prices to charge various applications, which can potentially lead to a revenue increase, as dictated by traditional economic theory.

This inspires us to propose an *"application-driven pricing approach"*. Formally, an ISP offers a number of different service classes, and each CP can choose a particular class to deliver the data for its application. The ISP guarantees the quality of service and determines a unit usage price for each service class. This quality guarantee can be of multiple dimensions, e.g., throughput, delay, packet drop rate, and the like.[1] To fulfill quality guarantee, the ISP needs to determine the amount of resources (e.g., bandwidth capacity) allocated to each service class. Content providers, on the other hand, are free to choose which class to use depending on the requirement of their applications.

Beyond the difficulty in determining the optimal choices of CPs and ISPs, it is important to answer whether this proposal can bring a "*significant*" revenue gain to ISPs, since only in this case ISPs will have incentives to make such change. However, there is no simple answer to this question. On the one hand, the heterogeneity of applications indicates that a multi-class pricing may be beneficial; on the other hand, this approach divides applications into various categories, reducing the utilization of statistical multiplexing of delivering the packets, which may reduce the revenue.

Our proposal can be regarded as a new implementation of differentiated pricing, besides a number of well-studied forms, e.g., congestion pricing [16], Paris Metro Pricing [22], DiffServ [3], and so on. Although different in objectives and methods, existing works share some technical similarity with our proposal, and they seem to be reporting something negative: Despite the fact that differentiated pricing is more profitable than single-class pricing under certain conditions, many of them [6, 13, 28, 35] show numerically that this gain is not significant, making it hard to be realized in practice. We will show completely different conclusions: The revenue gain can be *significant* under practical scenarios, making it meaningful for ISPs to make a change. This is because we apply a more accurate model and take important factors into consideration. Our contributions are:

[1] In general, QoS (quality of service) guarantee can be difficult over the Internet. However, there is a trend that the majority of contents will be directly transmitting between access ISPs and CPs without going through multiple ASes [7], so QoS can be controlled by this access ISP. In the remainder of this article, we refer to this access ISP as ISP for short.

- We develop a model to capture the heterogeneity of applications from multiple perspectives, i.e., the reservation price, the volume of traffic usage, and the quality requirement. We model how an ISP and CPs make decisions and capture the interactions between them.
- We formulate a revenue maximization framework to determine the optimal choices of the ISP and CPs.
- We analyze the impact of (1) heterogeneity of applications, (2) multiplexing, and (3) the system capability.
- We show that our proposal can improve the revenue of the ISP significantly (by as high as over 20%), especially when (1) the ISP's capacity is limited, (2) the applications are neutral to multiplexing, (3) the number of CPs with strict quality requirement is moderately small, and (4) CPs' reservation prices are not very low.

This article is organized as follows: In Section 2, we discuss related work. In Section 3, we present the system model, the design of application-driven pricing scheme, and an optimization framework to capture the ISP's decision. In Section 4, we analyze the optimal pricing problem of the ISP. In Section 5, we extend our model to capture the quality requirement heterogeneity. In Section 6, we study the revenue improvement quantitatively. In Section 7, we discuss net neutrality, service classes selection, and dynamic demand. In Section 9, we conclude.

2 RELATED WORK

We first summarize the difference of our approach with four typical *differentiated pricing approaches*.

- *Congestion Pricing*. Congestion pricing charges users a high price when the congestion occurs in data transmission. It reduces congestion because a high price discourages the usage. MacKie-Mason and Varian [16] showed that the congestion pricing scheme maximizes net social benefits. Paschalidis and Tsitsiklis [23] obtained some conditions under which the static pricing scheme achieves similarly good performance as the congestion pricing scheme. Henderson et al. [12] surveyed several possible approaches to implement the congestion pricing scheme. All these three works fall into the "single class setting", i.e., an ISP either provides one service class or provides multiple service classes without competition among service classes. Different from them, our approach considers multiple service classes with such competition. Furthermore, it is technically non-trivial to extend congestion pricing to this competitive multiple service class setting.
- *Paris Metro Pricing (PMP)*. PMP partitions a resource (e.g., a train) into several classes (e.g., ordinary and premium compartments) and controls their service quality via prices, such that a highly priced class will be with a high QoS, as fewer users will choose it. Odlyzko [22] claimed PMP as the simplest solution to service differentiation. Gibbens et al. [10] showed that PMP is not viable in a competitive market with multiple ISPs. Jain et al. [13] showed that PMP improves the revenue for a monopoly ISP under the capacity sharing metric, while Ros and Tuffin [27] showed an opposite result under the queuing latency metric. Shakkottai et al. [28] obtained sufficient conditions under which a single class achieves similar performance as PMP. Chau et al. [6] concluded that the viability of PMP for a monopoly ISP depends on how users react to congestion externality. Ma [14] also studied the viability of PMP under a more realistic model. Zhou et al. [35] considered the scenario that the ISP provides an infinite number of service classes with service qualities in a continuous domain, and applied an optimal control framework to derive the optimal prices for each service class. Different from PMP which is content-agnostic, our approach depends on the nature of applications so that requests can be better satisfied based on their specific requirement. In addition,

existing works either concluded theoretically the viability of PMP, or showed numerically that the revenue gain of multi-class pricing is not significant [6, 13, 28, 35], making it hard to realize in practice. We show completely different conclusions: The revenue gain can be *significant* under practical scenarios. This is because our approach and analysis capture *important aspects* of application heterogeneity, as we will see later.

- *Diffserv.* Blake et al. [3] proposed the Diffserv architecture, which provides multiple service classes over IP networks and marks packets to use a certain service class. Marbach [17, 18] proposed priority classes to charge packets, under which a user paying higher can have a higher priority to send its packets. Shu and Varaiya [29] proposed an auction-based approach to price different service classes. Wang and Schulzrinne [31] proposed a dynamic pricing scheme, which adjusts the price based on the usage, the level of service, and congestion. Unlike Diffserv, which takes the price as the only exogenous parameter and quality, can only be endogenous, we take QoS as an exogenous parameter. Diffserv offers options to increase the priority in sending packets, but it does not provide service quality guarantees, while our approach provides such guarantees. Furthermore, our approach provides a reasonable isolation for applications, which differentiates them based on the requirement on service quality. This isolation opens some potentials for better control design of the system.
- *Premium Peering.* Premium peering is a newly established business relationship where access ISPs charge CPs for the premium quality of services beyond the best effort, which is in fact a new form of differentiated service. In particular, Courcoubetis et al. [7] used a Nash bargaining approach to determine the premium peering prices. One interesting insight is that per service peering is a promising approach to resolve the interconnection tussles between ISPs and CPs, which shares a similar idea with our per-application pricing. Wang et al. [32] uncovered conditions under which the Internet access providers should offer CPs the option of paid peering. The premium peering approach only provides service quality guarantees for the superior class, leaving the inferior class without any service quality guarantee, while our approach provides service quality guarantees for all service classes. Unlike the work [7] which determines prices only (via Nash bargaining), we use an optimization framework to determine the optimal price, quality guarantee, and capacity for each service class. Different from the work [32] which studies the viability of paid peering, we focus on study when multi-class pricing can bring significant revenue gain.

Finally, we discuss some works that study the pricing problem in Internet applications, but do not fall into any one of the above four categories. Bhargava and Sun [2] proposed a multi-class pricing scheme with statistical QoS guarantees. Their model does not capture the congestion externality, while our model captures this important factor. Nault and Zimmerman [21] proposed a two-class pricing scheme. They assumed that there is no congestion externality in the superior service class and thus the service quality can be guaranteed regardless of the traffic volume. The service quality of the inferior class is affected by the congestion externality. Our model is more general in that the congestion externality exists in every service class. Wang et al. [30] studied optimal two-sided pricing under congestion externality. Their work aims to optimize pricing decisions from the ISP's perspective. They focused on single class pricing and derived a number of theoretical characterizations on the optimal pricing decisions.

3 MODEL

In this section, we set up a general model on the interactions between one ISP and a set of CPs. In particular, we model how CPs make decisions for content delivery services, and how the ISP

determines the number of service classes, as well as the price, capacity, and quality guarantee for each service class.

3.1 The Marketplace and Decision Spaces

As we have mentioned, the Internet market is a two-sided market where ISPs stand in between CPs and end-users. Accordingly, the pricing scheme of an ISP consists of two parts: How to charge CPs and how to charge end-users. In this article, we focus on the first perspective. This is because our proposal is application centric, which leads to a major change on the charging schemes to CPs. Usually CPs have limited choices on which ISPs to connect to; in other words, the ISPs often form an oligopoly market. As a first study, in this article, we consider only *one ISP* and focus on the interactions between this ISP and a set of CPs. This ISP announces to provide a set $\mathcal{N} \triangleq \{1, \ldots, N\}$ of service classes, each class $i \in \mathcal{N}$ with a quality guarantee q_i and correspondingly a *per unit usage price* p_i. Namely, this multi-class pricing is a usage-based pricing scheme. Each CP may choose any service class for its content delivery[2] depending on the requirement of its application service. We will model such choices later. We consider that the ISP uses the allocated capacity exclusively for each service class due to the following reasons. First, each service class has a quality guarantee, no matter it is a superior class or an inferior class. If we allow multiplexing, the service quality of some service classes may not be guaranteed. Second, allocating capacity exclusively for each service class also provides an isolation approach for applications, which differentiates them based on the requirement on service quality. This isolation opens some potentials for better control design of the system.

In general, a service quality guarantee might be multi-dimensional (e.g., throughput, delay, packet loss rate, etc.). In this article, we consider a one-dimensional conceptual quality guarantee measured by congestion level [15]. This is not only for mathematical tractability, but also because congestion is one of the most important features of the Internet. We will show later that this simplified model can already reveal a number of fundamental understandings. A larger value of q_i indicates a higher level of congestion, and thus a lower quality. If the ISP promises a quality guarantee of q_i, it means the congestion level is at most q_i.

The ISP has a total bandwidth capacity of C and it devotes $c_i \triangleq Ck_i$ to each service class i, where $\mathbf{k} \triangleq (k_i : i \in \mathcal{N}) \in \mathcal{K}$ and $\mathcal{K} \triangleq \{\mathbf{k} \in [0,1]^N : \sum_{i \in \mathcal{N}} k_i = 1\}$. We denote $\mathbf{p} \triangleq (p_i : i \in \mathcal{N}) \in \mathbb{R}_+^N$ and $\mathbf{q} \triangleq (q_i : i \in \mathcal{N}) \in \mathbb{R}_+^N$ as the vectors of prices and congestion levels for each service class respectively. Without loss of generality, we label service classes from inferior to superior, i.e., $q_1 > q_2 > \cdots > q_N$.

To summarize, the decision space for any CP is which *service class* to choose for its content delivery. The decision space for the ISP consists of the choices on the *quality guarantee*, *price* and *capacity* for each service class.

3.2 Content Providers' Choice Model

A particular CP k has a reservation price (or the willingness-to-pay) $v_k \in \mathcal{V} \triangleq [0, \hat{v}]$, where $\hat{v}$ is the maximum possible value for v_k. This v_k reflects the maximal unit usage price that a CP is willing to pay for its content delivery. CPs may have different reservation prices, for example, a video streaming usually has a higher reservation price than a file downloading service. If the ISP proposes a price higher than that, the content provider will not use this class. Once a CP k has chosen a service class with a unit usage price p and a congestion level q, we define its utility of using this service as its surplus (i.e., the difference between the actual fee it pays to the ISP, and

[2]We assume that each CP only operates one service. For CPs that provide multiple services, we can simply treat them as multiple individual CPs.

its willing-to-pay amount), in the following form:

$$u_k(p,q) \triangleq (v_k - p)\hat{d}_k \rho_k(q),$$

where $\hat{d}_k \rho_k(q)$ denotes the users' demand under congestion q, which is the maximum possible demand $\hat{d}_k$ under no congestion (i.e., $q=0$), multiplied by a discount function $\rho_k(q) \in [0,1]$. The discount function $\rho_k(q)$ models CP k's quality requirement, i.e., it will reduce its total usage of a content delivery service when the quality is not satisfactory, captured by the following assumption:

Assumption 1. *$\rho_k(q) : \mathbb{R}_+ \to [0,1]$ is continuous and decreasing in q. In addition, $\rho_k(0) = 1$ and $\lim_{q\to\infty} \rho_k(q) = 0$.*

To simplify the presentation, we define a dummy service class 0 with price $p_0 = 0$ and congestion $q_0 = +\infty$ to model the choice that a CP decides to opt out of all service classes. Formally, a CP k's choice is to select a service class i_k for its data delivery that maximizes its own utility, i.e.,

$$i_k \in \arg\max_{j \in \mathcal{N}\cup\{0\}} u_k(p_j, q_j).$$

When there is a tie, i.e., $|\arg\max_{j\in\mathcal{N}\cup\{0\}} u_k(p_j,q_j)| > 1$, the CP can simply break it by choosing the service class with the lowest price.

To summarize, we model the heterogeneity of applications from three perspectives: the reservation prices v_k, the maximum demand $\hat{d}_k$, and the quality requirement $\rho_k(q)$. In what follows, we will first assume a homogeneous discount function for all CPs, denoted by $\rho(q)$, which can be interpreted as the average quality requirement over the whole CP population. Under this simplification, we reduce the extent of heterogeneity so as to deliver rigorous analytical results. *We will study the general setting with heterogeneous discount functions in Section 5 and Section 6.*

3.3 Demand in Each Service Class

Now we investigate how many CPs will choose a particular service class and what is the total demand they will incur. This is important because it has a major impact on the workload of the ISP and thus influence the ISP's decision. Recall that CP's applications differ in nature, and this is represented by the reservation price v_k and unconstrained demand $\hat{d}_k$ for each CP. Due to this heterogeneity, different CPs may choose different service classes and this choice only depends on v_k as $\hat{d}_k$ is a linear coefficient of a CP's utility. Given $q_i < q_j$, we define

$$v_{ij} \triangleq \frac{p_i\rho(q_i) - p_j\rho(q_j)}{\rho(q_i) - \rho(q_j)}$$

as the *boundary price*, such that a CP k prefers service class i over j (i.e., $u_k(p_i,q_i) > u_k(p_j,q_j)$) if and only if $v_k > v_{ij}$. In other words, service classes i and j are *indifferent* to any CP whose reservation price is equal to v_{ij}. Let us denote $v_i^{max} \triangleq \max\{v_{ij} : 0 \le j < i\}$, i.e., the maximal boundary price when service class i is compared with all inferior classes. Content providers with reservation prices greater than v_i^{max} will not choose service classes inferior to class i. Then a CP k chooses class N if $v_k > v_N^{max}$. We derive the *market segment* for class N as $(\min\{\hat{v}, v_N^{max}\}, \hat{v}]$. Similarly, we can derive the market segment for class i, denoted by $(V_{i-1}, V_i]$, as

$$V_i = \begin{cases} \hat{v}, & \text{if } i = N, \\ \min\{V_{i+1}, v_{i+1}^{max}\}, & \text{otherwise.} \end{cases} \tag{1}$$

We consider a continuum spectrum of CPs. Let $F_V(v)$ denote the mass of the maximum demand for all CPs having the reservation price less or equal to v. We assume a well-formed function F_V, such that F_V is twice differentiable and increasing in $v \in \mathcal{V}$. This continuum model is quite

common in network service pricing works [6, 14]. In the following lemma, we derive the actual demand of each service class.

Lemma 1. *Given* $\mathbf{p}$ *and* $\mathbf{q}$*, the aggregated demand for any non-dummy service class* i *is* $d_i = D_i(\mathbf{p}, \mathbf{q})$*, where* $D_i(\mathbf{p}, \mathbf{q})$ *is defined as*

$$D_i(\mathbf{p}, \mathbf{q}) \triangleq [F_V(V_i) - F_V(V_{i-1})]\,\rho(q_i), \quad \forall i \in \mathcal{N}.$$

Further, $D_i(\mathbf{p}, \mathbf{q})$ *is continuous, non-increasing in* p_i (q_i) *and non-decreasing in* p_j (q_j) *for all* $j \neq i$*.*

Proofs to lemmas and theorems can be found in Section 8. Lemma 1 states that the demand for each service class is non-increasing in the price and congestion level of this class, and non-decreasing in the price and congestion level of any other class.

3.4 Steady-State of the Market

Recall that the ISP's decision space includes $\mathbf{p}, \mathbf{q}$, and $\mathbf{k}$. These decision variables are inter-dependent in the sense that any variable may have an impact on others, and in turn impacts itself in a feedback loop. This is really a dynamic system, and we are interested in defining the "steady state" of this market via the concept of equilibrium.

We first consider a particular service class with a fixed volume of aggregated demand of CPs (denoted by d_i), and a fixed amount of bandwidth capacity dedicated to this class (denoted by c_i). It is natural to assume that the congestion q_i is uniquely determined by d_i and c_i. Formally, we define a congestion function $Q(d_i, c_i)$ such that $q_i = Q(d_i, c_i)$.

Assumption 2. $Q(d_i, c_i) : \mathbb{R}_+^2 \to \mathbb{R}_+$ *is continuous, increasing in* d_i*, decreasing in* c_i *and satisfies* $Q(d_i, 0) = +\infty$*.*

We next define the "implied demand", which will be useful later to define the market equilibrium.

Definition 1. Given a congestion level q_i and a capacity Ck_i, the implied demand of class i is denoted by $\Delta(q_i, k_i) : \mathbb{R}_{\geq 0}^2 \to \mathbb{R}_{\geq 0}$. If $Q(0, Ck_i) \leq q_i$, then $\Delta(q_i, k_i)$ is the unique demand which satisfies $q_i = Q(\Delta(q_i, k_i), Ck_i)$, otherwise $\Delta(q_i, k_i) = 0$.

If this *implied demand* equals the *actual demand*, then, from the definition, we can see that the actual congestion level equals the ISP's commitment, so CPs have no incentives to change their choices, and the system reaches a steady state.

Definition 2. A tuple $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ is an equilibrium, if and only if it satisfies the following equations:

$$D_i(\mathbf{p}, \mathbf{q}) = \Delta(q_i, k_i), \quad \forall i \in \mathcal{N}.$$

We can see that in order to achieve an equilibrium, a tuple $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ has only two degrees of freedom, i.e., if any two vectors are properly given, then the third vector can be uniquely determined at the equilibrium. In what follows, we state how to determine $\mathbf{p}$ if $\mathbf{q}$ and $\mathbf{k}$ are given.

Lemma 2. *Suppose* $\mathbf{q}$ *and* $\mathbf{k}$ *are given such that* $\Delta(q_i, k_i) > 0$ *for any* $k_i > 0$*. Then a tuple* $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ *is an equilibrium if and only if* $p_i = P_i(\mathbf{q}, \mathbf{k})$ *holds for all* $i \in \mathcal{N}$*, where*

$$P_i(\mathbf{q}, \mathbf{k}) \triangleq \frac{1}{\rho(q_i)} \sum_{j=1}^{i} \Big[\rho(q_j) - \rho(q_{j-1})\Big] \bar{F}_V^{-1}(\theta_j), \tag{2}$$

where $\theta_j \triangleq \sum_{\ell=j}^{N} \Delta(q_\ell, k_\ell)/\rho(q_\ell)$ *defines the aggregate implied maximum demand for all the service classes having congestion no worse than class* j*, and* $\bar{F}_V^{-1}$ *is the inverse of* $\bar{F}_V(v) \triangleq F_V(\hat{v}) - F_V(v)$*.*

This lemma guides the ISP to set the appropriate price to commit its quality guarantee for each service class. Later, we will use this relationship at the equilibrium to analyze the optimal pricing

strategies of the ISP. Note that, in this lemma, we require that $\mathbf{q}$ and $\mathbf{k}$ are given such that $\Delta(q_i, k_i) > 0$ for all $k_i > 0$. It requires that for any given capacity c_i, the ISP proposes an appropriate quality guarantee so as to attract a positive demand. Otherwise, this class does not bring any revenue.[3] From an ISP's perspective, it is equivalent to setting $k_i = 0$ for this class. Thus, in our further analysis, we can safely remove this trivial case from our consideration.

3.5 ISP's Decision Model

The ISP's objective is to select $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ to maximize its own revenue. We assume a volume-based charging scheme, and define the ISP's revenue as the total income charged from CPs for delivering their content in an equilibrium $(\mathbf{p}, \mathbf{q}, \mathbf{k})$:

$$R(\mathbf{p}, \mathbf{q}, \mathbf{k}) \triangleq \sum_{i \in \mathcal{N}} p_i D_i(\mathbf{p}, \mathbf{q}).$$

Thus, the ISP's decision problem can be formulated as:

Problem 1. *ISP's revenue maximization at equilibrium.*

$$\begin{aligned} \underset{\mathbf{p},\mathbf{q},\mathbf{k}}{\textit{maximize}} \quad & R(\mathbf{p}, \mathbf{q}, \mathbf{k}) \\ \textit{subject to} \quad & D_i(\mathbf{p}, \mathbf{q}) = \Delta(q_i, k_i), && \forall i \in \mathcal{N}, \\ & \mathbf{p} \in \mathbb{R}_+^N, \mathbf{q} \in \mathbb{R}_+^N, \mathbf{k} \in \mathcal{K}. \end{aligned}$$

Our further analysis, i.e., whether and to what extent differentiated pricing can bring higher revenue for the ISP, will be based on this neat form. Note that this formulation implicitly assumes that when making its own decision, the ISP takes the CPs' decisions into consideration, which has been reflected in the equilibrium constraint. In fact, we are using a Stackelberg game model to capture the interactions between the ISP and CPs: The ISP is the first mover who decides $(\mathbf{p}, \mathbf{q}, \mathbf{k})$, and CPs are second movers who decide which service class to select. Our above formulation will lead to the solution (or the Stackelberg equilibrium) of this game.

As we will see later, analyzing this problem is non-trivial. We consider a finite N, because in reality an ISP can only provide a finite number of service classes. The ISP can allocate zero capacity to some service classes so as to close them. We call those with positive capacities as "active classes", which are our real focus. Later, we will use $\mathcal{N}_a$ to represent the set of active classes and denote $N_a = |\mathcal{N}_a|$. We can see that based on this definition, analyzing Problem 1 can be divided into the following two steps:

Step 1: Determine the optimal N_a.

Step 2: Put N_a classes into $\mathcal{N}_a$ and solve:

Problem 2. *Revenue maximization for active service classes.*

$$\begin{aligned} \underset{\mathbf{p},\mathbf{q},\mathbf{k}}{\textit{maximize}} \quad & R(\mathbf{p}, \mathbf{q}, \mathbf{k}) \\ \textit{subject to} \quad & D_i(\mathbf{p}, \mathbf{q}) = \Delta(q_i, k_i), && \forall i \in \mathcal{N}_a, \\ & k_i > 0, \Delta(q_i, k_i) > 0, && \forall i \in \mathcal{N}_a, \\ & \mathbf{p} \in \mathbb{R}_+^{N_a}, \mathbf{q} \in \mathbb{R}_+^{N_a}, \mathbf{k} \in \mathcal{K}. \end{aligned}$$

The above two steps not only clearly state whether multi-class pricing or single-class pricing is preferred, but also ensure that any optimal solution to Problem 2 is an interior point. This is because each active class has a positive capacity (i.e., $k_i \in (0, 1], \forall i \in \mathcal{N}_a$), so a solution can be optimal

[3] A typical example is the queuing latency metric $Q(d_i, c_i) = 1/(c_i - d_i)$ [6], where $Q(d_i, c_i) = +\infty$ for all $d_i \geq c_i$. Suppose q_i is given to be $1/c_i$. The corresponding d_i has to be 0.

only if the revenue for each active class is positive, which implies that the price and quality guarantee must lie in the interior domain (i.e., $p_i \in (0, \hat{v}), q_i \in (0, \infty), \forall i \in \mathcal{N}_a$). Ensuring that an optimal solution is an interior point will make our further analysis easier.

Remark. Numerically, the ISP can solve Problem 1 to get the optimal price, quality guarantee and capacity for each service class and solving Problem 1 does not rely on Equation (2). In other words, for the numerical study purpose, the ISP does not need to follow the two-step process: (1) determine the number of active service classes; (2) determine the optimal price, quality guarantee, and capacity for each active service class via solving Problem 2. This two-step process and Equation (2) are mainly developed to facilitate the theoretical analysis of Problem 1.

3.6 Physical Interpretation

The formulation can already reveal some fundamental factors that impact the decision of the ISP.

- *Application Heterogeneity.* Our model characterizes the application heterogeneity in terms of the reservation price and maximum demand. If they are of high diversity, then D_i can be easily differentiated by setting proper prices. As economic theory dictates, this can bring a revenue gain.
- *Statistical Multiplexing.* The demand function D_i is constrained by the congestion function at the equilibrium. Thus, if the congestion function exhibits the statistical multiplexing property (i.e., the traffic that a service class can support increases more rapidly than its own capacity), the total demand $\sum_i D_i(\mathbf{p}, \mathbf{q})$ may decrease when N_a is large, so the total revenue may drop for the multi-class pricing scheme.
- *ISP Capability.* The maximal number of classes allowed (N) and total capacity (C) impact the value of D_i through the implied demand function. Their effect on the preference/objection of multi-class pricing is even more vague: once such parameters change, the revenues of the single- and multi-class pricing change in the same trend, so it is difficult to tell which one will be preferred.

To summarize, there are really conflicting (i.e., heterogeneity vs. multiplexing) and unclear (i.e., ISP capability) factors that make it difficult to determine whether the multi-class pricing can be more profitable. In later sections, we will decouple them and show their individual impacts.

4 OPTIMAL PRICING

In this section, we first analyze the number of active service classes, followed by a formal characterization on the optimal pricing strategy of these active classes. We then characterize properties of the maximal revenue of the ISP asymptotically.

4.1 Number of Active Service Classes

The number of active service classes is important because it indicates whether and how the multi-class pricing scheme will be implemented. For example, if there are $N_a < N$ active classes, the ISP does not really need to implement N service classes; instead, implementing N_a classes is just enough. In particular, if there is only one active service class, it indicates the traditional single-class pricing scheme is already the best strategy. We will show that this number N_a is deeply impacted by the property of the congestion function Q. Recall that we have defined the implied demand function $\Delta(q_i, k_i)$. Let us define three families of congestion functions based on $\Delta(q_i, k_i)$.

Definition 3. A congestion function Q is multiplexing-preferred, or multiplexing-neutral, or anti-multiplexing, if the implied demand function satisfies that $\Delta(q_i, k_i - k_i') + \Delta(q_i, k_i')$ is less

than, or equal to, or larger than $\Delta(q_i, k_i)$, where $k_i' < k_i$, $\Delta(q_i, k_i - k_i') > 0$, $\Delta(q_i, k_i') > 0$ and $\Delta(q_i, k_i) > 0$.

This division is based on whether an ISP can support more demand by merging two service classes of the same quality guarantee into one and keeping the quality unchanged. If yes, we call Q "multiplexing-preferred". If the demand that can be supported remains unchanged, it means that applications are not sensitive to multiplexing and we call Q "multiplexing-neutral". On the other hand, if the demand reduces after merging, we call Q "anti-multiplexing".

We consider two typically used congestion functions and relate them to the above definition.

- *Queuing Latency Based.* The congestion function based on the queuing latency model $Q(d_i, c_i) = 1/(c_i - d_i)$ is commonly used in prior works, e.g., [6], [14], [27], and [30], where $Q(d_i, c_i) = +\infty$ for all $d_i \geq c_i$. One can have $\Delta(q_i, k_i) = \max\{Ck_i - 1/q_i, 0\}$. It can be simply verified to be multiplexing-preferred. This congestion function captures the total expected waiting time (i.e., queuing plus service time) in an M/M/1 queue setting with service rate c_i and arrival rate d_i. It is suitable to capture the congestion in UDP-based applications, where packets queue up and are served in a FIFO pattern.
- *Capacity Sharing Based.* Another commonly used congestion function (e.g., in [6], [13], [14], [30], [32], and [35]) is based on the capacity sharing model $Q(d_i, c_i) = d_i/c_i$. One can have $\Delta(q_i, k_i) = Ck_iq_i$. It is simple to verify it as neutral to multiplexing. This metric models that various applications share the capacity of network systems, and the ratio of bandwidth allocated to each application is proportional to the individual demand. It is suitable to capture the congestion in TCP-based applications, where the data backs up when there is a congestion, and are supposed to be delivered finally.

Though the two aforementioned families of congestion functions were proposed to model network congestion externality decades ago [13, 27], they are still commonly used in recent works [6, 14, 30, 32, 35]. The reason is that they are mathematical simple yet have strong physical meanings. Thus, we will only focus on the two aforementioned families of congestion functions. We will not discuss anti-multiplexing congestion functions since we don't find any practical ones.

4.1.1 Multiplexing-Preferred Congestion Functions. We derive sufficient conditions under which the best strategy of the ISP is to implement only one active service class.

Theorem 1. *Suppose the congestion function Q satisfies*

$$\Delta(q_i, k_i - k_i') + \Delta(q_i', k_i') < \Delta(q_i, k_i), \tag{3}$$

where $\Delta(q_i, k_i - k_i') > 0$, $\Delta(q_i', k_i') > 0$, and $\Delta(q_i, k_i) > 0$. If $F_V(v)$ is convex, the ISP maximizes its revenue only if there is only one active service class, i.e., $N_a = 1$.

Theorem 1 states the conditions under which the best decision for the ISP is the conventional single-class pricing.[4] Note that in here, our requirement is even stricter than the definition of multiplexing-preferred functions. Condition (3) states that the demand will reduce if the ISP partitions a service class into two subclasses, even if the quality guarantee of one subclass is allowed to change. In other words, Condition (3) specifies a *subset* of multiplexing-preferred congestion functions and in fact, they are *strongly* multiplexing-preferred.

One may feel that our requirement in (3) is too strict to be realistic, but it is really practical. In fact, the congestion function based on queuing latency model satisfies this condition, because $\Delta(q_i, k_i - k_i') + \Delta(q_i', k_i') = Ck_i - 1/q_i - 1/q_i' < \Delta(q_i, k_i)$, and thus we have the following corollary.

[4] A similar idea was reported in [6]. However, the definition, model, and proof method differ from our model.

Corollary 1. *If $Q(d_i, c_i) = 1/(c_i - d_i)$ and $F_V(v)$ is convex, the ISP maximizes its revenue by implementing only one active service class.*

Requiring a convex $F_V(v)$ implies that a significant number of CPs have high reservation prices, since $F_V(v)$ is convex if and only if the density function $f_V(v) \triangleq \frac{dF_V(v)}{dv}$ is non-decreasing. This assumption is mainly for the mathematical tractability in our proof. Our simulation results show that this might not be a necessary condition, though we lack a proof. In particular, in our simulation, we consider $Q(d_i, c_i) = 1/(c_i - d_i)$ as well as a concave and an S-shaped instance of F, i.e.,

$$F_V(v) = v^{\alpha}, \qquad \forall \alpha \in (0,1), v \in [0,1],$$

$$F_V(v) = \frac{1}{Z}\frac{1}{1+\exp(-\alpha(v-0.5))}, \quad \forall v \in [0,1], \alpha \in (0,10], Z = \int_0^1 \frac{1}{1+\exp(-\alpha(v-0.5))}dv.$$

We find that Theorem 1 still holds.

4.1.2 Multiplexing-Neutral Congestion Functions. Now let us discuss what is the number of active service classes when the congestion function is neutral to multiplexing.

Theorem 2. *If the congestion function Q is neutral to multiplexing, the ISP maximizes its revenue only if the number of active service classes is $N_a = N$.*

Theorem 2 states that if the congestion function is neutral to multiplexing, then the ISP will simply activate all service classes so as to increase its revenue. Recall that a capacity-sharing-based congestion function is multiplexing-neutral. Thus, we have the following corollary:

Corollary 2. *If $Q(d_i, c_i) = d_i/c_i$, the ISP maximizes its revenue by implementing N service classes.*

Please note that our conclusion does not depend on the value of N. Thus, if N is large, the ISP will provide a large number of service classes. In other words, if N can be chosen, theoretically the ISP may want to provide an infinite number of active service classes to increase its revenue. However, we will show in next subsection that the increase of revenue is bounded with respect to the number of active service classes. In addition, offering a large number of service classes often means a heavy operational cost. We therefore conclude that it is only realistic for the ISP to activate a limited number of service classes. The above observations are similar to two previous works in the nonlinear pricing literature: (1) Wilson [33] proved that increasing the number of tariff options always increases the revenue; (2) Miravete [19] found that a finite number of tariff options is offered in practice, due to the cost in offering tariff options such as commercialization cost.

Physical Interpretation. Recall that there are conflicting (heterogeneity vs. multiplexing) and unclear factors (ISP capability parameters) for adopting or rejecting multi-class pricing. In this section, we decouple them and only focus on the impact of multiplexing effect. Our results show that when the congestion function exhibits a "strong" multiplexing property, the ISP does not need to consider multi-class pricing; when the congestion function is neutral to multiplexing, the ISP always prefer multi-class pricing. This result does not depend on any particular shape of the function $F_V(v)$ or the ISP's capability parameters. We do not conduct further analysis on other congestion functions, since the commonly applied ones have already been captured.

We have taken an initial step in understanding the optimal pricing strategy of the ISP, i.e., whether the ISP should provide single or multiple active service classes. Given the number of active service classes, we next analyze the optimal pricing decisions $(\mathbf{p}, \mathbf{q}, \mathbf{k})$.

4.2 Optimal Pricing for Active Service Classes

Note that we have determined the number of active service classes N_a, now we proceed to analyze Problem 2. Recall that the optimal solution must be an interior solution to Problem 2, since each active class has a non-zero capacity. By using the KKT conditions, we have the following theorem:

Theorem 3. *Consider Problem 2. If* $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ *is an optimal solution, for any active class* $i \in \mathcal{N}_a$*, we have the following:*

$$\begin{cases} \left[\theta_{i+1}\bar{F}_V^{-1}(\theta_{i+1}) - \theta_i \bar{F}_V^{-1}(\theta_i)\right] \dfrac{1}{B_i} = \sum_{j=1}^{i} A_j, & \forall i \in \mathcal{N}_a, \\ C_i \sum_{j=1}^{i} A_j = C_1 A_1, & \forall i \in \mathcal{N}_a, \end{cases}$$

where A_j, B_i*, and* C_i *are defined as:*

$$A_j \triangleq [\rho(q_j) - \rho(q_{j-1})] \frac{\partial \theta_j \bar{F}_V^{-1}(\theta_j)}{\partial \theta_j}, \quad B_i \triangleq \frac{1}{\rho'(q_i)} \frac{\partial \Delta(q_i, k_i)/\rho(q_i)}{\partial q_i}, \quad C_i \triangleq \frac{1}{\rho(q_i)} \frac{\partial \Delta(q_i, k_i)}{\partial k_i}.$$

Theorem 3 states a necessary condition for $\mathbf{q}$ and $\mathbf{k}$ to be optimal, in forms of an array of equations. Once we have $\mathbf{q}$ and $\mathbf{k}$, the optimal price $\mathbf{p}$ can be calculated base on Equation (2). One can solve these equations to obtain the optimal $\mathbf{q}$ and $\mathbf{k}$. In some scenarios, it might be computationally more efficient than solving the original optimization problem directly. Furthermore, we can utilize it to calculate the optimal single-class quality guarantee.

Corollary 3. *Consider single-class pricing* $N = 1$*. Problem 1 has at least one optimal solution. The optimal single-class quality guarantee* q *satisfies* $\epsilon_{\bar{F}} = \epsilon_\rho / \epsilon_\Delta - 1$ *where*

$$\epsilon_{\bar{F}} \triangleq \frac{d\bar{F}_V}{dp} \frac{p}{\bar{F}_V}, \quad \epsilon_\rho \triangleq \frac{d\rho}{dq} \frac{q}{\rho}, \quad \epsilon_\Delta \triangleq \frac{\partial \Delta(q,1)}{\partial q} \frac{q}{\Delta(q,1)}.$$

If $\epsilon_{\bar{F}}$ *is decreasing in* p*,* ϵ_ρ *is non-increasing in* q *and* ϵ_Δ *is non-increasing in* q*, then the optimal single-class quality guarantee* q *uniquely satisfies* $\epsilon_{\bar{F}} = \epsilon_\rho / \epsilon_\Delta - 1$*.*

Corollary 3 states sufficient conditions under which the optimal single-class quality guarantee q is unique. Under these conditions, one can design bi-section algorithms to locate the optimal single-class quality guarantee q.

4.3 Asymptotic Analysis of Maximum Revenue

We have revealed some insights to calculate the optimal decision variables for the multi- and single-class pricing. It is mathematically intractable to analyze the revenue improvement in general since we lack closed-form optimal decision variables. Asymptotically, we can analyze the maximum revenue without closed-form optimal decision variables, and through this we draw some insights on the revenue improvement. In particular, we characterize the impact of the ISP's capability parameters, i.e., C and N, on the maximum revenue.

Theorem 4 (Impact of Capacity). *The ISP's maximal revenue increases in* C *and approaches* $\max_p p\bar{F}(p)$ *as* $C \to \infty$*.*

Theorem 4 states that the ISP can increase its revenue by expanding its capacity, but this increment is upper bounded. Note that this result holds regardless of the value of N. It implies that the single-class pricing scheme approaches this limit as well, so the multi-class pricing will have a marginal revenue gain over the single-class pricing when the capacity is sufficiently large. Thus, the ISP does not need to consider multi-class pricing when the capacity is sufficiently large.

Theorem 5 (Impact of Number of Service Classes). *The maximal revenue of the ISP is non-decreasing with respect to N and is upper bounded by $\max_p p\bar{F}_V(p)$ as $N \to \infty$.*

Recall that when Q is multiplexing-neutral, the ISP can always increase revenue by providing one more service class. Theorem 5 states that this increase is upper bounded when N approaches infinity, or the revenue improvement has a *diminishing return to scale* effect. This statement is similar to a previous work by Wilson [33], where Wilson proved that the revenue improvement by increasing the number of tariff options decreases in the number of tariff options. Taking the operating cost into the consideration, the ISP will provide a limited number of service classes in reality.

We have revealed some insights on revenue improvement in asymptotic cases. What remains unknown are: (1) In general, how ISP's capability parameters impact revenue improvement; and (2) how application heterogeneity impacts revenue improvement. Let's proceed to investigate them.

5 EXTENSIONS

Recall that we pointed out three aspects of application heterogeneity, i.e., the reservation price, the maximum demand and the quality requirement. Our model thus far captures the first two aspects, showing that multi-class pricing can always improve revenue when applications are neutral to multiplexing. In order to quantitatively study the revenue improvement in practical scenarios, now let us generalize our model to capture the heterogeneity in quality requirement.

5.1 Model Extensions

Our model has assumed that all CPs have the same discount function $\rho(q)$. Now we extend it to capture the heterogeneity of applications in quality requirement. Let the discount function for CP k be denoted by $\rho_k(q) \triangleq \rho(w_k, q)$, where $w_k \in \mathcal{W} \triangleq [0, \hat{w}]$ models CP k's quality requirement, and $\hat{w}$ denotes the maximum possible value for w_k. A larger value of w_k means that a CP has a stricter quality requirement. We have the following assumption.

Assumption 3. *$\rho(w_k, q) : \mathbb{R}_+^2 \to [0, 1]$ is continuous and decreasing in w_k. In addition, $\rho(0, q) = 1$ for all q.*

We define the aggregate maximum demand for all CPs having reservation prices no larger than v and quality requirements no larger than w as

$$F_{V,W}(v, w) \triangleq \sum_k \mathbf{1}_{\{v_k \le v,\, w_k \le w\}} \hat{d}_k. \tag{4}$$

Accordingly, we generalize ISP's decision problem (i.e., Problem (1)) via generalizing the demand function $D_i(\mathbf{p}, \mathbf{q})$ as

$$D_i(\mathbf{p}, \mathbf{q}) = \int_0^{\hat{w}} \int_{V_{i-1}(w)}^{V_i(w)} f_{V,W}(v, w)\rho(w, q_i) dv dw, \tag{5}$$

where $f_{V,W}(v, w) \triangleq \frac{\partial^2 F_{V,W}(v,w)}{\partial v \partial w}$ and $V_i(w)$ is obtained by replacing $\rho(q_i)$ in Equation (1) with $\rho(w, q_i)$. By using the KKT conditions, we characterize the optimal solution for the generalized ISP decision problem as follows:

Theorem 6. *A tuple $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ maximizes the ISP's revenue, only if it satisfies*

$$\begin{cases} \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial p_i} \Big/ \dfrac{\partial G}{\partial p_i} = \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial p_{i+1}} \Big/ \dfrac{\partial G}{\partial p_{i+1}}, & i = 1, \ldots, N_a - 1, \\ \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial p_i} \Big/ \dfrac{\partial G}{\partial p_i} = \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial q_i} \Big/ \left(\dfrac{\partial G}{\partial q_i} - H_i\right), & i \in \mathcal{N}_a, \end{cases}$$

where G and H_i are defined as

$$G \triangleq \sum_{j\in\mathcal{N}_a} D_j(\mathbf{p},\mathbf{q}) \Big/ \frac{\partial\Delta(q_j,k_j)}{\partial k_j}, \quad H_i \triangleq D_i(\mathbf{p},\mathbf{q}) \frac{\partial^2\Delta(q_i,k_i)}{\partial q_i \partial k_i} \Big/ \left[\frac{\partial\Delta(q_i,k_i)}{\partial k_i}\right]^2 - \frac{\partial\Delta(q_i,k_i)}{\partial q_i} \Big/ \frac{\partial\Delta(q_i,k_i)}{\partial k_i},$$

and $D_i(\mathbf{p},\mathbf{q})$ is derived in Equation (5).

One can solve these equations to obtain the optimal decision variables. In some scenarios, it might be computationally more efficient than solving the original optimization problem directly (e.g., by using the gradient methods.).

It is difficult to gain further insights from an abstract form of the CPs distribution $F_{V,W}(v,w)$. Thus, we seek a reasonable instantiation of such distribution so that we can later show the revenue gain of multi-class pricing. Before that, let us first state some scaling properties, based on which we can realize such instantiation.

5.2 Scaling Properties

We first present some scaling properties to simplify the presentation of the instantiation on $F_{V,W}(v,w)$.

Theorem 7. *Suppose $(\mathbf{p},\mathbf{q},\mathbf{k})$ is an optimal solution. If we scale $\hat{v}$ by $\xi > 0$ such that $\tilde{v} \in [0,\xi\hat{v}]$, and $\tilde{F}_{V,W}(\tilde{v},w) = F_{V,W}(\tilde{v}/\xi,w)$, then after scaling, $(\xi\mathbf{p},\mathbf{q},\mathbf{k})$ is an optimal solution and the maximum revenue scales by ξ.*

Theorem 7 states that as we scale the reservation price linearly, the maximal revenue also scales linearly. Thus, we can normalize the reservation price to be in $\mathcal{V} = [0,1]$.

Let us consider a quintessential form $\rho(w,q) = e^{-wq}$ of demand discount function, which has been used to study congestion externality of product market [26] and has been applied to network service pricing [14, 15].

Theorem 8. *Suppose $\rho(w,q) = e^{-wq}$, $Q(d_i,c_i) = d_i/c_i$ and $(\mathbf{p},\mathbf{q},\mathbf{k})$ is an optimal solution. If we scale C and $\hat{w}$ such that $\widetilde{C} = \xi C$, $\tilde{w} \in [0,\xi\hat{w}]$ and $\widetilde{F}_{V,W}(v,\tilde{w}) = F_{V,W}(v,\tilde{w}/\xi)$, then after scaling, $(\mathbf{p},\mathbf{q}/\xi,\mathbf{k})$ is an optimal solution with the maximal revenue unchanged.*

Theorem 8 states if we scale the capacity and quality requirement linearly at the same rate, the optimal prices, capacities, and the maximal revenue keep unchanged, and the optimal quality guarantee scales. Thus, we normalize the quality requirement to be in $\mathcal{W} = [0,1]$.

Theorem 9. *Let $Q(d_i,c_i) = d_i/c_i$ and $(\mathbf{p},\mathbf{q},\mathbf{k})$ is an optimal solution. If we scale C and $F_{V,W}(v,w)$ such that $\widetilde{C} = \xi C$ and $\widetilde{F}_{V,W}(v,w) = \xi F_{V,W}(v,w)$, then after scaling, $(\mathbf{p},\mathbf{q},\mathbf{k})$ is an optimal solution and the maximal revenue scales by ξ.*

Theorem 9 states that when C and $F_{V,W}(v,w)$ scale linearly at the same rate, the optimal pricing scheme does not change and the maximal revenue scales linearly. Thus, we normalize the CPs distribution such that $F_{V,W}(\hat{v},\hat{w}) = 1$.

To summarize, we normalize $F_{V,W}(v,w)$ such that $(v,w) \in [0,1]^2$ and $F_{V,W}(1,1) = 1$ without loss of any generality. In other words, $F_{V,W}(v,w)$ can be treated as a probability distribution over $(v,w) \in [0,1]^2$. Based on it, we next instantiate the CPs distribution $F_{V,W}(v,w)$.

5.3 Instantiation on the Distribution of CPs

Now let us consider how to construct a reasonable distribution of CPs via function $F_{V,W}(v,w)$. It is important to capture the heterogeneity in (1) reservation prices and (2) quality requirements, as well as their *correlations*. Since $F_{V,W}(v,w)$ is a probability distribution over $(v,w) \in [0,1]^2$, we use

the marginal distribution $F_V(v)$ to model price heterogeneity, the marginal distribution $F_W(w)$ to model quality requirement heterogeneity, and the Kendall tau correlation coefficient [1] to quantify the *correlation*:

$$\tau \triangleq \mathbb{P}[(V-\tilde{V})(W-\tilde{W})>0]-\mathbb{P}[(V-\tilde{V})(W-\tilde{W})<0],$$

where $(V,W),(\tilde{V},\tilde{W}) \sim F_{V,W}(v,w)$. The value of $\tau \in [-1,1]$, where $\tau>0$ ($\tau<0$) models positive (negative) dependency. For example, consider the following $F_{V,W}(v,w)$ [1]:

$$F_{V,W}(v,w) = \exp\left[-\left((-\alpha \ln v)^{\theta} + (-\beta \ln w)^{\theta}\right)^{\frac{1}{\theta}}\right], \tag{6}$$

where $(v,w) \in [0,1]^2$ and $\theta \geq 1$. It characterizes the price heterogeneity, quality requirement heterogeneity and dependency via parameters α, β and θ, respectively. First, the marginal price distribution is $F_V(v) = v^{\alpha}, v \in [0,1]$. Thus, α models the distribution of CPs with respect to the reservation price. For example, $\alpha = 1$ corresponds to a uniform distribution of CPs' reservation prices, and $\alpha < 1$ ($\alpha > 1$) corresponds the distribution leaning towards low (high) values. Second, the marginal quality requirement distribution is $F_W(w) = w^{\beta}, w \in [0,1]$. Last, the Kendall tau correlation coefficient is $\tau = 1 - \frac{1}{\theta}$ with $\theta \geq 1$. It captures the full regime of positive dependency, i.e., a high (or low) reservation price tends to accompany a strict (or loose) quality requirement.

6 REVENUE IMPROVEMENT

We aim at a quantitative study on the amount of revenue gain of multi-class pricing over single-class pricing, so we can understand under what scenarios our proposal is meaningful to implement. *Note that in this section, we consider the general setting with heterogeneous discount functions.*

6.1 Experiment Settings

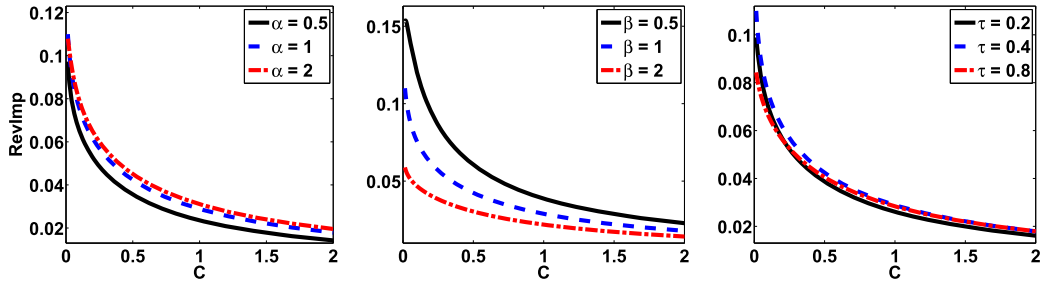
We have revealed the impact of multiplexing. Since the congestion function based on the queuing model always prefers single-class pricing, we will not discuss it in this section. Instead, we will focus on the congestion function based on the capacity sharing model, i.e., $Q(d_i, c_i) = d_i/c_i$. As we have stated, this form represents TCP-like applications, which is the majority of today's Internet traffic. We have shown that the revenue gain has a diminishing return to scale effect with respect to the number of service classes. Thus, in this section, we will focus on the revenue gain of two-class pricing over single-class pricing. Formally, we define the revenue improvement as

$$\text{RevImp} \triangleq \frac{R_2^* - R_1^*}{R_1^*}, \tag{7}$$

where R_1^* and R_2^* denote the maximal revenue under single- and two-class pricing, respectively.

We consider one ISP and a continuum spectrum of CPs with heterogeneous requirement on service quality, i.e., they are described by the model in Section 5. The aggregate maximum demand $F_{V,W}(v,w)$ satisfies Equation (4) and the demand function $D_i(\mathbf{p},\mathbf{q})$ satisfies Equation (5). We extend Problem 1 to determine the optimal price, quality guarantee, capacity and revenue, where we replace the demand function in Problem 1 with Equation (5). By the scaling properties in Section 5.2, we normalize the aggregate maximum demand $F_{V,W}(v,w)$ such that $(v,w) \in [0,1]^2$ and $F_{V,W}(1,1) = 1$. Furthermore, we consider the instance of aggregate maximum demand $F_{V,W}$ expressed in Equation (6). As the aggregate maximum demand from all the CPs is normalized to $F_{V,W}(1,1) = 1$, we consider a capacity varies from 0 to 2, i.e., $C \in [0,2]$, where $C = 2$ models that the capacity is abundant.

Though captured asymptotically, we are still unaware how the capacity will impact the revenue improvement in general cases. Also, we have not discussed the impact of application heterogeneity.



(a) $\beta = 1, \tau = 0.4$ (b) $\alpha = 1, \tau = 0.4$ (c) $\alpha = 1, \beta = 1$

Fig. 1. Impact of C on revenue improvement.

Therefore, our focus in this section is to investigate the impact of capacity and the heterogeneity of applications on the revenue improvement.

6.2 Impact of Capacity

We first examine the impact of capacity on revenue improvement. In this study, we choose typical cases of $F_{V,W}(v, w)$ to characterize application heterogeneity:

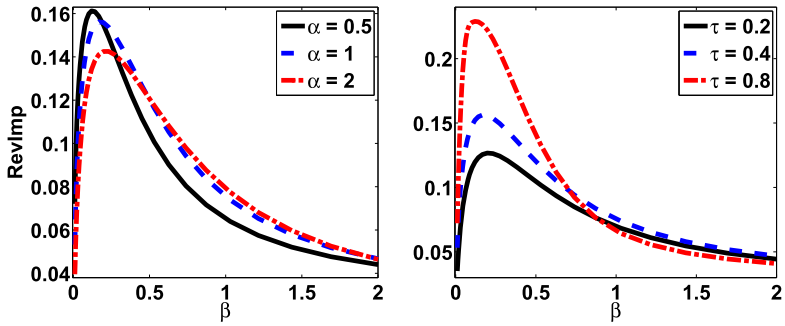
- The distribution of the reservation price parameter α leans towards a low value ($\alpha = 0.5$), or it is uniform ($\alpha = 1$), or it leans towards a high value ($\alpha = 2$);
- The distribution of the quality requirement parameter β leans towards a low value ($\beta = 0.5$), or it is uniform ($\beta = 1$), or it leans towards a high value ($\beta = 2$); and
- The reservation price and quality requirement are positively dependent and the degree of dependency is low ($\tau = 0.2$), or medium ($\tau = 0.4$), or high ($\tau = 0.8$).

Figure 1 plots the revenue improvement when the capacity varies. From Figure 1, we observe more than 10% revenue improvement when the capacity is small. When C increases, RevImp reduces till around 2%. This implies that when the capacity is limited (or abundant), the improvement is significant (or marginal). This is natural since when C becomes small (or large), providing high QoS is difficult (or easy) and it is meaningful (or meaningless) to do differentiation. Also, note that in reality, due to increasing demand, bandwidth has always been a competing resource, so there is a high chance to gain a significant revenue improvement. The impact of application heterogeneity (i.e., α, β and τ) on revenue improvement is non-trivial, which is our remaining focus.

Lessons learned. The revenue improvement can be significant (i.e., more than 10% in general, and as high as over 15%) when the capacity is limited.

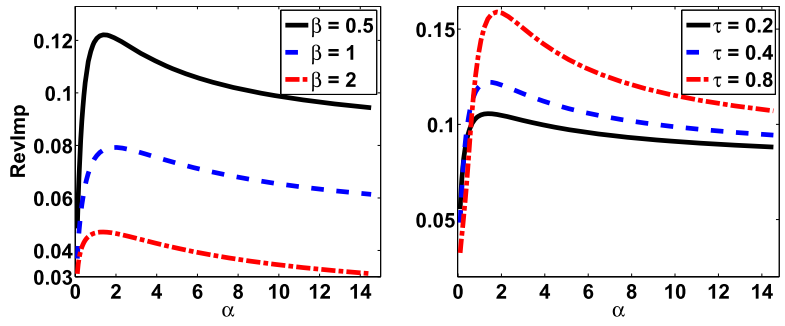
6.3 Impact of Quality Requirement Heterogeneity

Now we fix $C = 0.1$ (i.e., the capacity is limited) and investigate the impact of application heterogeneity (α, β, and τ). We first study the impact of β, which captures the distribution of CPs with respect to the quality requirement. Figure 2 plots the revenue improvement when β varies. We can observe that when β is moderately small, the revenue improvement can be more than 20%. This indicates that when a moderately small number of CPs have strict quality requirements, the ISP can improve the revenue significantly by two-class pricing. This is because strict (or loose) quality requirements often indicate high (or low) reservation prices, which can be satisfied by a superior (or inferior) service class. Given that the superior class charges at a higher price, when the number of CPs in the superior class is moderately small, the total revenue from the two classes are comparable and thus both are significant. If changing to a single-class pricing, the ISP deems to lose



(a) $\tau = 0.2, C = 0.1$ (b) $\alpha = 1, C = 0.1$

Fig. 2. Impact of β on revenue improvement.



(a) $\tau = 0.4, C = 0.1$ (b) $\beta = 0.5, C = 0.1$

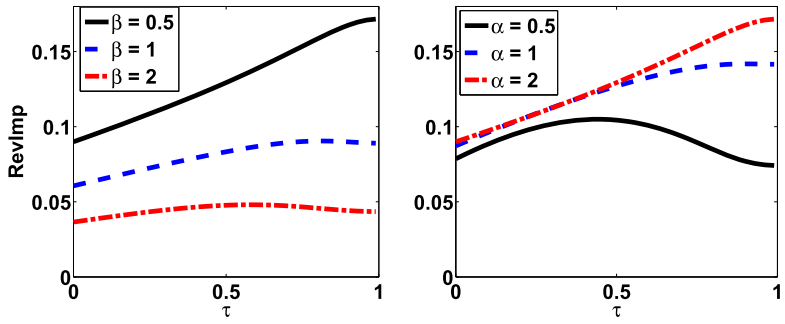
Fig. 3. Impact of α on revenue improvement.

revenue from either of the original two classes. When β is very small or very large, the revenue improvement becomes marginal, i.e., around 4%. It implies that when CPs' quality requirement distribution is concentrated, the diversity of quality requirement is low, so single-class pricing is sufficient. We also observe that the price heterogeneity and dependency have a non-trivial impact on revenue improvement. We next investigate their roles.

Lessons learned: The revenue improvement is significant (i.e., more than 10% in general, and as high as over 20%) when a moderately small number of CPs have strict requirement on the service quality.

6.4 Impact of Reservation Price Heterogeneity

Now let us investigate the impact of α, which captures the distribution of CPs with respect to the reservation price. Figure 3 plots the revenue improvement when α varies. Similar to the previous subsection, we fix $C = 0.1$ and β at proper values. From Figure 3, we observe more than 10% revenue improvement when: (1) α is not very small; and (2) β is small, which further verifies our observations in the last subsection. This implies that when most CPs' reservation prices are not very low, the revenue improvement can be significant. We also observe that when α is very small, the revenue improvement is marginal, i.e., around 4%. This implies that when the reservation price distribution leans closely towards a low value, the ISP does not need to consider multi-class pricing. When the α increases from a small value, the revenue improvement increases. This is because the diversity of reservation price increases as α increases. However, when α further increases, the revenue improvement starts decreasing. This is because when α becomes large, the



(a) $\alpha = 2, C = 0.1$ (b) $\beta = 0.5, C = 0.1$

Fig. 4. Impact of τ on revenue improvement.

reservation price distribution leans closely towards a high value, or its diversity becomes low again. Last, the correlation coefficient τ has a significant impact on the improvement, which we will further investigate.

Lessons learned. The revenue improvement is significant (i.e., more than 10% in general, and as high as over 15%) when CPs' reservation prices are not very low.

6.5 Impact of Dependency

Now we investigate the impact of τ, which quantifies the dependency between the reservation price and the quality requirement. Figure 4 plots the revenue improvement when τ varies. Similarly we fix $C = 0.1$ and set α, β at proper values. From Figure 4, we observe more than 10% revenue improvement when τ is not very small. This implies that the revenue improvement can be significant when the dependency between the reservation price and quality requirement is not very weak. It can be even more than 15% when τ is large, i.e., they are strongly positively correlated. This is because the ISP can use a superior (or inferior) class to satisfy CPs with high (or low) reservation prices and strict (or loose) quality requirements. The revenue improvement curves exhibits a complicated trend, some of which are monotone while others are not. One reason is that when τ varies, the revenues of both single-class and two-class pricing vary in the same trend, and their ratio may exhibit a non-monotone feature.

Lessons learned: The revenue improvement is significant (i.e., more than 10% in general, and as high as over 15%) when the dependency between reservation price and quality requirement is not very weak.

6.6 Impact of the Number of Service Classes

Now we investigate the number of services classes N. Table 1 shows the optimal revenue and revenue improvement when the number of service classes varies from 1 to 4. Similarly, we fix $c = 0.1$ and set α, β, τ at proper values. From Table 1, we observe that the optimal revenue increases in the number of service classes N. This further validates our analytical results. Increasing the number of service classes from $N = 1$ to $N = 2$ can increase the revenue by as high as 15.85%, i.e., significant revenue improvement. The revenue improvement is around 3% when the number of service classes increases from $N = 2$ to $N = 3$, and around 1% when the number of service classes increases from $N = 3$ to $N = 4$. In other words, further increasing the number of service classes beyond three only brings marginal revenue improvement. This implies a strong diminishing return of revenue improvement, i.e., the revenue improvement by increasing the number of service classes decreases rapidly in the number of service classes. This observation is similar to two previous works in the

Table 1. Impact of the Number of Service Classes on Revenue Improvement ($C = 0.1$)

Parameters			**Revenue**				**Revenue improvement**		
α	β	τ	R_1^*	R_2^*	R_3^*	R_4^*	$\frac{R_2^*-R_1^*}{R_1^*}$	$\frac{R_3^*-R_2^*}{R_2^*}$	$\frac{R_4^*-R_3^*}{R_3^*}$
0.5	0.5	0.4	0.007168	0.007918	0.008134	0.008223	10.46%	2.73%	1.08%
1	0.5	0.4	0.011051	0.012379	0.012761	0.012918	12.02%	3.09%	1.23%
2	0.5	0.4	0.016313	0.018282	0.018851	0.019086	12.07%	3.11%	1.25%
2	**0.5**	0.4	0.016313	0.018282	0.018851	0.019086	12.07%	3.11%	1.25%
2	**1**	0.4	0.012931	0.013955	0.014256	0.014383	7.92%	2.16%	0.89%
2	**2**	0.4	0.011116	0.011632	0.01178	0.011846	4.65%	1.29%	0.54%
2	0.5	**0.2**	0.017620	0.019466	0.020002	0.020227	10.47%	2.76%	1.12%
2	0.5	**0.4**	0.016313	0.018282	0.018851	0.019086	12.07%	3.11%	1.25%
2	0.5	**0.8**	0.014502	0.016801	0.017514	0.017813	15.85%	4.24%	1.71%

nonlinear pricing literature: (1) Wilson [33] theoretically proved that the revenue improvement by increasing the number of tariff options decreases rapidly in the number of tariff options; (2) Courty and Paglieo [8] found empirical evidences for this theoretical result of Wilson in the context of offering multi-class pricing for concert tickets. Even though our context and model are different from theirs, our idea is similar to theirs, i.e., using multi-class pricing to extract surplus from consumers (i.e., CPs in our context) so as to improve the revenue. For our setting, the intuitive explanation for the strong diminishing return of revenue improvement is the following. Revenue improvement by increasing the number of service classes is achieved by using more service classes to extract more surplus from CPs. Note that the total amount of surplus from CPs is finite. As the number of service classes increases, there is less and less surplus from CPs left. Furthermore, providing one more service classes becomes less and less "attractive" to CPs as there may already exists some comparable alternative service classes, making the newly added service class less and less powerful in extracting surplus from CPs. Lastly, the revenue improvement is always positive, which implies an ISP should provide as many classes of service as possible. We will show in Section 7 that this strong diminishing return of revenue improvement together with the cost of offering service classes leads to a finite number of service classes being provided in reality.

Lessons learned. The revenue improvement is significant (i.e., more than 10% in general, and as high as over 15%) when the number of service classes increases from one to two. Further increasing the number of service classes beyond three only brings marginal revenue improvement.

7 DISCUSSION

Selecting the Number of Services Classes. Consider the capacity sharing congestion model. Even though increasing the number of service classes can always increase the revenue (Theorem 2), in practice an ISP may offer a small number of service classes, due the cost in offering service classes. Formally, let $\tilde{C}(N) \in \mathbb{R}_+$ denote the cost of offering N classes of service. Let $R_N^* - \tilde{C}(N)$ denote the profit of offering N classes of service. The ISP selects N to maximize the profit. Table 2 shows the impact of N on the ISP's profit. One can observe that when the cost function is $\tilde{C}(N) = 0.0015(N-0.5)^2$, i.e., convex, the profit decreases as N increases from 1 to 4. This implies that the ISP should offer one service class in order to attain the maximum profit. Similarly, when the cost function is $\tilde{C}(N) = 0.0015(N-0.5)$, i.e., linear, the ISP should offer two service classes. When the cost function is $\tilde{C}(N) = 0.0015(N-0.5)^{0.5}$, i.e., concave, the ISP should offer three service classes.

Comparing the Revenue with Other Pricing Schemes. One may suggest to compare the revenue of our pricing scheme with the pricing schemes summarized in Section 2. However, we find it

Table 2. Impact of N on the ISP's Profit ($\alpha = 2, \beta = 0.5, \tau = 0.8, c = 0.1$)

	Profit			
Cost function	$R_1^* - \tilde{C}(1)$	$R_2^* - \tilde{C}(2)$	$R_3^* - \tilde{C}(3)$	$R_4^* - \tilde{C}(4)$
Convex: $\tilde{C}(N) = 0.0015(N - 0.5)^2$	0.0141	0.0134	0.0081	−0.0006
Linear: $\tilde{C}(N) = 0.0015(N - 0.5)$	0.0138	0.0146	0.0138	0.0126
Concave: $\tilde{C}(N) = 0.0015(N - 0.5)^{0.5}$	0.0034	0.0150	0.0151	0.0150

difficult to conduct a fair and meaningful comparison of revenues, as these pricing schemes consider different scenarios or models. Some of the reasons are the following:

- As discussed in Section 2, the congestion pricing falls into the "single class setting". It adjusts the price according to the real-time congestion dynamically. In other words, it sets price to control the evolving dynamics of the congestion. However, our pricing scheme sets pricing parameters based on the congestion equilibrium. In other words, our pricing scheme controls the equilibrium of congestion. Adapting the congestion pricing into our model, we obtain a single class pricing scheme. This adaption loses the key feature of the congestion pricing, i.e., controlling the evolving dynamics of the congestion, leading to unfair comparisons.
- As discussed in Section 2, the Diffserv considers a model at the packet routing level, where setting the pricing parameters requires the adjustment of priority in sending packets. However, our model is at the traffic volume level, i.e., abstracts a CP's total traffic packets as a real number. Under this abstraction our model isolates the traffic volumes based on their service quality requirement and it does not involve priority adjustment in sending packets. Thus, it is difficult to adapt the Differv into our model and select the appropriate parameters for it.
- As discussed in Section 2, the premium peering provides two classes of service, i.e., a superior class and an inferior class. It provides service quality guarantees for the superior class, leaving the inferior class without any service quality guarantee. However, our pricing scheme aims to provide service quality guarantees for all service classes. Thus, it is difficult to select appropriate parameters for premium peering to achieve a fair comparison.
- As discussed in Section 2, existing works either concluded theoretically the viability of PMP, or showed numerically that the revenue gain of multi-class pricing is not significant, making it hard to realize in practice. We show completely different conclusions: The revenue gain can be significant under practical scenarios. This is because our approach and analysis capture important aspects of application heterogeneity.
- In the end of Section 2, we summarize several multi-class pricing schemes that provide service quality guarantees. These pricing schemes do not consider congestion externality. Our approach considers a more realistic setting of multi-class pricing under congestion externality. Adapting these pricing schemes into model leads to the failure of providing service quality guarantees.

Net Neutrality. The motivation of net neutrality is to provide fair and efficient operation environment for all applications, promoting a healthy growth of the ecosystem. Note that in practice, there is no consensus on the exact definition of neutrality means [24]. One common (and strict) definition that we find is "Network neutrality refers to the principle of an agnostic network, that is, one which does not discriminate against the content which travels across it or the applications or

hardware which engage with and connect to it [5]." It means that the Internet should be regarded as a utility like electricity, and that every one has the right to use it equally. There are arguments on whether net neutrality is the right way to achieving the motivation. For example, Crowcroft says that "we never had network neutrality in the past, and I do not believe we should engineer for it in the future either" [9]. Misra states that "network neutrality issue is really about economics rather than freedom or promoting/stifling innovation [20]."

In terms of real implementation, rules differ across countries. In U.S., the debate of whether/how to perform regulation is still in debate. FCC regulated that "data transmission is transparent; ISPs cannot enforce unreasonable interference or disadvantage end customers", etc., but did not regulate specialized services. In practice, FCC only imposed regulation rules on the ISP-user side, but did not impose rules on the ISP-CP side. One of the reason is that there have been many complicated and hidden commercial contracts between ISPs and CPs (people call it "premium peering"), making it extremely difficult, if not impossible, to request publicity of all these commercial secrets, not to mention to remove all of them. Even for the regulation on the ISP-user side, there have been many years of debate and and recently, FCC has lifted all regulations regarding network neutrality. In Europe, competition between fixed and mobile broadband providers works much better than in the US, so there is less concern about neutrality.

Our proposal may not be considered as fully satisfying the strict net neutrality definition in [5]. However, it aligns with the FCC's previous regulations (which we think was a very strict regulation implementation in the real market). Most importantly, in regard of the initial motivation of net neutrality, we would like to emphasize that if implemented, our pricing scheme does not only increase ISPs' revenue, but it is beneficial to the whole ecosystem. CPs do not need to construct their own networks any more, but they can buy specialized services with lower costs. As applications are delivered with quality guarantee, users' quality of experience will be improved. In summary, this can potentially increase the social welfare of the entire Internet ecosystem, which *is* the original motivation of the net neutrality.

Dynamic Demand. Once the behaviors of CPs (e.g., d_k) change, one can input the new parameter regarding CPs (e.g., d_k) into our framework to re-configure the price accordingly. This provides a simple solution to setting that the behaviors of CPs change over time. A more formal method to handle this setting is the dynamic pricing, which is one of the future works on this article.

8 PROOFS OF LEMMAS AND THEOREMS

Proof of Lemma 1. Note that if a CP k's reservation price v_k is in segment $[V_{i-1}, V_i)$, it will choose service class i. Hence, the aggregate maximum demand for service class i is $F_V(V_i) - F_V(V_{i-1})$. We obtain $D_i(\mathbf{p}, \mathbf{q})$ by noting that all CPs choosing class i have the same discount function $\rho(q_i)$. Observe that both V_i and V_{i-1} are continuous in p_i (q_i) and $p_j(q_j)$. Hence, $D_i(\mathbf{p}, \mathbf{q})$ is continuous in p_i (q_i) and $p_j(q_j)$. When p_i increases, $u_k(p_i, q_i)$ decreases. Based on the choice model, we know the amount of users choosing class i will not increase. Hence, $D_i(\mathbf{p}, \mathbf{q})$ is non-increasing in p_i. Similarly, we reach the monotonicity conclusion. □

Proof of Lemma 2. The tuple $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ is an equilibrium if and only if it satisfies $F_V(V_i) - F_V(V_{i-1}) = \Delta(q_i, k_i)/\rho(q_i), \forall i \in \mathcal{N}$. The remaining task is to solve this equation array to obtain the price p_i. Note that $F_V(V_i) - F_V(V_{i-1}) = \Delta(q_i, k_i)/\rho(q_i)$ implies that $\sum_{\ell=i}^{N}[F_V(V_i) - F_V(V_{i-1})] = \sum_{\ell=i}^{N} \Delta(q_i, k_i)/\rho(q_i)$ Then, it follows that $F_V(\hat{v}) - F_V(V_{i-1}) = \sum_{\ell=i}^{N} \Delta(q_\ell, k_\ell)/\rho(q_\ell)$. By solving this equation, we have $V_{i-1} = \bar{F}_V^{-1}(\sum_{\ell=i}^{N} \Delta(q_\ell, k_\ell)/\rho(q_\ell)) = \bar{F}_V^{-1}(\theta_i)$. Note that $\mathbf{q}, \mathbf{k}$ are given such that $\Delta(q_i, k_i) > 0$. Hence, we have that $V_i > V_{i-1}$. Recall that the utility function $u_k(p, q)$ is linear with respect to the reservation price v_k. Then it follows that $V_i = [p_{i+1}\rho(q_{i+1}) - p_i\rho(q_i)]/[\rho(q_{i+1}) - \rho(q_i)]$. Hence, we have $p_i\rho(q_i) - p_{i-1}\rho(q_{i-1}) = [\rho(q_i) - \rho(q_{i-1})]\bar{F}_V^{-1}(\theta_i)$, which

implies that $\sum_{j=1}^{i}[p_i\rho(q_i) - p_{i-1}\rho(q_{i-1})] = \sum_{j=1}^{i}[\rho(q_i) - \rho(q_{i-1})]\bar{F}_V^{-1}(\theta_i)$. Observe that $p_0\rho(q_0) = 0$. Hence, we have $p_i\rho(q_i) = \sum_{j=1}^{i}[\rho(q_j) - \rho(q_{j-1})]\bar{F}_V^{-1}(\theta_j)$. Thus, we complete the proof. □

Proof of Theorem 1. In order to achieve the conclusion, we only need to show that, for each $N_a > 1$, $R(\mathbf{p}, \mathbf{q}, \mathbf{k}) < R_1^*$. To facilitate the analysis, we apply Lemma 2 to transform Problem 2 into the following form:

$$\begin{aligned} \underset{\mathbf{q},\mathbf{k}}{\text{maximize}} \quad & \gamma(\mathbf{q}, \mathbf{k}) \triangleq \sum_{i \in \mathcal{N}_a} [\rho(q_i) - \rho(q_{i-1})]\theta_i \bar{F}_V^{-1}(\theta_i) \\ \text{subject to} \quad & \theta_1 \le F_V(\hat{v}), \mathbf{q} \in \mathbb{R}_+^{N_a}, \mathbf{k} \in \mathcal{K}. \end{aligned}$$

Based on this formulation, our objective is to show that $\gamma(\mathbf{q}, \mathbf{k}) < R_1^*, \forall N_a > 1$. For simplicity of presentation, we define $\widetilde{R}(x) \triangleq x\bar{F}_V^{-1}(x)$. Let us first use the fact that the function $\widetilde{R}(x)$ is concave in x. We shall prove this fact later. We can derive an upper bound for the revenue as follows:

$$\begin{aligned} \gamma(\mathbf{q}, \mathbf{k}) &= \sum_{i \in \mathcal{N}_a} [\rho(q_i) - \rho(q_{i-1})]\widetilde{R}(\theta_i) = \rho(q_{N_a}) \sum_{i \in \mathcal{N}_a} \frac{\rho(q_i) - \rho(q_{i-1})}{\rho(q_{N_a})} \widetilde{R}(\theta_i) \\ &\le \rho(q_{N_a})\widetilde{R}\left(\sum_{i \in \mathcal{N}_a} \frac{\rho(q_i) - \rho(q_{i-1})}{\rho(q_{N_a})} \theta_i \right), \end{aligned} \tag{8}$$

where the last inequality holds because $\widetilde{R}(x)$ is concave in x, $(\rho(q_i) - \rho(q_{i-1}))/\rho(q_{N_a}) \ge 0$ and $\sum_{i \in \mathcal{N}_a}(\rho(q_i) - \rho(q_{i-1}))/\rho(q_{N_a}) = (\rho(q_{N_a}) - \rho(q_0))/\rho(q_{N_a}) = 1$. Note that for any given N_a, we have $\theta_i = \sum_{\ell=i}^{N_a} \Delta(q_\ell, k_\ell)/\rho(q_\ell)$. Then, it follows that

$$\begin{aligned} \sum_{i \in \mathcal{N}_a} \frac{\rho(q_i) - \rho(q_{i-1})}{\rho(q_{N_a})} \theta_i &= \theta_{N_a} + \sum_{i=1}^{N_a - 1} \frac{\rho(q_i)(\theta_i - \theta_{i+1})}{\rho(q_{N_a})} \\ &= \theta_{N_a} + \sum_{i=1}^{N_a - 1} \frac{\rho(q_i)}{\rho(q_{N_a})} \left(\sum_{\ell=i}^{N_a} \frac{\Delta(q_\ell, k_\ell)}{\rho(q_\ell)} - \sum_{\ell=i+1}^{N_a} \frac{\Delta(q_\ell, k_\ell)}{\rho(q_\ell)} \right) \\ &= \theta_{N_a} + \sum_{i=1}^{N_a - 1} \frac{\rho(q_i)}{\rho(q_{N_a})} \frac{\Delta(q_i, k_i)}{\rho(q_i)} = \frac{\sum_{i=1}^{N_a} \Delta(q_i, k_i)}{\rho(q_{N_a})} < \frac{\Delta(q_{N_a}, 1)}{\rho(q_{N_a})}, \end{aligned}$$

where the last inequality holds due to Condition (3). Note that the implied demand function $\Delta(q_i, k_i)$ increases in k_i. Hence, there exists a $k' < 1$ such that $\sum_{\ell=i}^{N_a} \Delta(q_i, k_i)/\rho(q_{N_a}) = \Delta(q_{N_a}, k')/\rho(q_{N_a})$. We can further derive the upper bound of $\gamma(\mathbf{q}, \mathbf{k})$ expressed in inequality (8) as

$$\gamma(\mathbf{q}, \mathbf{k}) \le \Delta(q_{N_a}, k')\bar{F}_V^{-1}\left(\frac{\Delta(q_{N_a}, k')}{\rho(q_{N_a})} \right)$$

The left side corresponds to the revenue of a single class pricing with capacity $k'C$ and quality guarantee q_{N_a}. Not that $k' < 1$, then we have

$$\gamma(\mathbf{q}, \mathbf{k}) \le \max_{q_{N_a}} \Delta(q_{N_a}, k')\bar{F}_V^{-1}\left(\frac{\Delta(q_{N_a}, k')}{\rho(q_{N_a})} \right) < \max_{q} \Delta(q, 1)\bar{F}_V^{-1}\left(\frac{\Delta(q, 1)}{\rho(q)} \right) = R_1^*,$$

where the last inequality holds because of the fact that the maximum revenue for single class pricing is increasing in capacity C (we will formally prove it in Theorem 4). This concludes the main result of this theorem.

Finally, let us show $\widetilde{R}(x) = x\bar{F}_V^{-1}(x)$ is concave in x. The second order derivative of $\widetilde{R}(x)$ is

$$\widetilde{R}''(x) = 2\frac{d\bar{F}_V^{-1}(x)}{dx} + x\frac{d^2\bar{F}_V^{-1}(x)}{dx^2}.$$

We can derive $\frac{d\bar{F}_V^{-1}(x)}{dx}$ as

$$\frac{d\bar{F}_V^{-1}(x)}{dx} = \frac{1}{\bar{F}_V'(\bar{F}_V^{-1}(x))} < 0,$$

where the last inequality holds because $\bar{F}_V(x)$ is decreasing in x. We can derive $\frac{d^2\bar{F}_V^{-1}(x)}{dx^2}$ as

$$\frac{d^2\bar{F}_V^{-1}(x)}{dx^2} = -\frac{\bar{F}_V''(\bar{F}_V^{-1}(x))}{(\bar{F}_V'(\bar{F}_V^{-1}(x)))^3} < 0,$$

where the last inequality holds because that $\bar{F}_V(x)$ is decreasing in x and concave in x (since $F_V(x)$ is convex in x). Hence, we have $\widetilde{R}''(x) < 0$, or $\widetilde{R}(x)$ is concave in x.

Now we complete the proof. □

Proof of Theorem 2. Our main idea is to show that the ISP can always increase its revenue by adding an inferior service class over the optimal class pricing for any given N_a. To facilitate the analysis, we use the equivalent formulation of Problem 2 used in the proof of Theorem 1. Note that in this formulation we only need to determine the optimal quality guarantee and the capacity for each active service class, as the optimal price $\mathbf{p}$ can be calculated base on them via Equation (2). Let $\mathbf{q}, \mathbf{k}$ denote the optimal quality guarantee and the capacity, respectively. Now suppose we divide the capacity of class 1, i.e., $k_1 C$ (which is the class with the lowest quality guarantee), to form two new service classes, where the first one is with quality guarantee q_1 and capacity $k_1 - k_a$, and the second one is with quality guarantee $q_a > q_1$ and capacity $k_a C < k_1 C$. The revenue after capacity dividing is

$$\widehat{\gamma}(q_a, k_a) \triangleq \sum_{i=2}^{N_a}[\rho(q_i) - \rho(q_{i-1})]\widetilde{R}(\theta_i) + \rho(q_a)\widetilde{R}(\theta_a) + [\rho(q_1) - \rho(q_a)]\widetilde{R}\left(\theta_2 + \frac{\Delta(q_1, k_1 - k_a)}{\rho(q_1)}\right),$$

where $\theta_i = \sum_{\ell=i}^{N}\Delta(q_\ell, k_\ell)/\rho(q_\ell),\ \forall i \geq 2,\ \theta_a = \theta_2 + \Delta(q_1, k_1 - k_a)/\rho(q_1) + \Delta(q_a, k_a)/\rho(q_a),$ and $\widetilde{R}(x) = x\bar{F}_V^{-1}(x)$. Then one can observe that $\widehat{\gamma}(q_a, 0) = \gamma(\mathbf{q}, \mathbf{k})$ is the maximum revenue before capacity dividing. We next show that, given $q_a > q_1$, there exists a $k_a \in (0, k_1)$ such that the revenue after capacity dividing become larger, i.e., $\widehat{\gamma}(q_a, 0) < \widehat{\gamma}(q_a, k_a)$. It suffices to show $\lim_{k_a \to 0^+}\frac{\partial\widehat{\gamma}(q_a, k_a)}{\partial k_a} > 0$. Now let us prove it.

Note that the congestion function is neutral to multiplexing, i.e., $\Delta(q_i, k_i - k_i') + \Delta(q_i, k_i') = \Delta(q_i, k_i)$. This implies that $\Delta(q_i, k_i)$ is a linear function of k_i. In particular, the implied demand function can be written in form of $\Delta(q_i, k_i) = Ck_i\zeta(q_i)$, where $\zeta(q_i) : \mathbb{R}^+ \to \mathbb{R}^+$ is a decreasing function of q_i. Then we have

$$\frac{\partial\widehat{\gamma}(q_a, k_a)}{\partial k_a} = \rho(q_a)\widetilde{R}'(\theta_a)\left(\frac{C\zeta(q_a)}{\rho(q_a)} - \frac{C\zeta(q_1)}{\rho(q_1)}\right) - (\rho(q_1) - \rho(q_a))\widetilde{R}'\left(\theta_2 + \frac{\Delta(q_1, k_1 - k_a)}{\rho(q_1)}\right)\frac{C\zeta(q_1)}{\rho(q_1)}.$$

Note that

$$\lim_{k_a \to 0^+}\theta_a = \lim_{k_a \to 0^+}\left[\theta_2 + \frac{\Delta(q_1, k_1 - k_a)}{\rho(q_1)} + \frac{\Delta(q_a, k_a)}{\rho(q_a)}\right] = \theta_2 + \frac{\Delta(q_1, k_1)}{\rho(q_1)} = \theta_1,$$

$$\lim_{k_a \to 0^+}\left[\theta_2 + \frac{\Delta(q_1, k_1 - k_a)}{\rho(q_1)}\right] = \theta_2 + \frac{\Delta(q_1, k_1)}{\rho(q_1)} = \theta_1.$$

Then it follows that

$$\lim_{k_a\to 0^+}\frac{\partial\widehat{\gamma}(q_a,k_a)}{\partial k_a}=\rho(q_a)\widetilde{R}'(\theta_1)\left(\frac{C\zeta(q_a)}{\rho(q_a)}-\frac{C\zeta(q_1)}{\rho(q_1)}\right)-[\rho(q_1)-\rho(q_a)]\widetilde{R}'(\theta_1)\frac{C\zeta(q_1)}{\rho(q_1)}$$
$$=C\widetilde{R}'(\theta_1)\left[\zeta(q_a)-\zeta(q_1)\right].$$

Observe that $\zeta(q_a)>\zeta(q_1)$ (since $q_a>q_1$). Hence, to show $\lim_{k_a\to 0^+}\frac{\partial\widehat{\gamma}(q_a,k_a)}{\partial k_a}>0$, we only need to show $\widetilde{R}'(\theta_1)>0$.

Note that $\mathbf{q},\mathbf{k}$ are optimal solutions. By applying the KKT conditions, we know that $\mathbf{q},\mathbf{k}$ must satisfy $\frac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial q_{N_a}}=0$. The physical intuition is that the ISP has no incentives the adjust the quality guarantee q_{N_a}. By expanding this equation, we have

$$\sum_{i\in\mathcal{N}_a}[\rho(q_i)-\rho(q_{i-1})]\widetilde{R}'(\theta_i)\frac{\partial\theta_i}{\partial q_{N_a}}+\rho'(q_{N_a})\widetilde{R}(\theta_a)=0.$$

Note that

$$\frac{\partial\theta_i}{\partial q_{N_a}}=\frac{\partial\Delta(q_{N_a},k_{N_a})/\rho(q_{N_a})}{\partial k_{N_a}}=C\frac{\zeta(q_{N_a})}{\rho(q_{N_a})}>0.$$

Hence, $\sum_{i\in\mathcal{N}_a}[\rho(q_i)-\rho(q_{i-1})]\widetilde{R}'(\theta_i)>0$. Similarly, by applying the KKT conditions, we know that $\mathbf{q},\mathbf{k}$ must satisfy the equation $\frac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial k_1}=\frac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial k_{N_a}}$ also. The physical intuition is that the ISP has no incentives to change the capacity allocation between class 1 and class N_a. By expanding this equation, we have

$$\rho(q_1)\widetilde{R}'(\theta_1)\frac{\partial\theta_1}{\partial k_1}=\sum_{i\in\mathcal{N}_a}[\rho(q_i)-\rho(q_{i-1})]\widetilde{R}'(\theta_i)\frac{\partial\theta_i}{\partial k_{N_a}}.$$

Note that

$$\frac{\partial\theta_i}{\partial k_{N_a}}=\frac{\partial\Delta(q_{N_a},k_{N_a})/\rho(q_{N_a})}{\partial k_{N_a}}=C\frac{\zeta(q_{N_a})}{\rho(q_{N_a})}.$$

Hence, the sign of $\widetilde{R}'(\theta_1)$ is the same as $\sum_{i\in\mathcal{N}_a}[\rho(q_i)-\rho(q_{i-1})]\widetilde{R}'(\theta_i)>0$. Thus, we complete the proof. □

Proof of Theorem 3. We prove this theorem by analyzing the KKT conditions on the equivalent formulation of Problem 2, which is used in the proof of Theorem 1. First we show that the constraint $\theta_1\le F_V(\hat{v})$ cannot hold in equality $\theta_1=F_V(\hat{v})$ in the optimal pricing. This is because that $\theta_1=F_V(\hat{v})$ implies that the price of class 1 is $p_1=0$. Intuitively, this cannot be optimal because the ISP invests a positive capacity k_1 while obtaining zero revenue from it (we will formally prove the maximum revenue increases with respect to the capacity in Theorem 4). Hence, it is only possible to have $\theta_1<F_V(\hat{v})$ under the optimal pricing. Based on it, by applying KKT conditions, we can see that $\mathbf{q},\mathbf{k}$ are optimal only if they satisfy

$$\begin{cases}\dfrac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial q_i}=0, & \forall i\in\mathcal{N}_a,\\ \dfrac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial k_i}=\dfrac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial k_1}, & \forall i\in\mathcal{N}_a.\end{cases}$$

The equation $\frac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial q_i}=0$ means that the ISP has no incentives to adjust the quality guarantee. By expanding it, we have

$$\sum_{j=1}^{i}[\rho(q_j)-\rho(q_{j-1})]\widetilde{R}'(\theta_j)\frac{\partial\Delta(q_i,k_i)/\rho(q_i)}{\partial q_i}+\rho'(q_i)\left[\widetilde{R}(\theta_i)-\widetilde{R}(\theta_{i+1})\right]=0.$$

Then, by rearranging terms, we have that $[\theta_{i+1}\bar{F}_V^{-1}(\theta_{i+1}) - \theta_i\bar{F}_V^{-1}(\theta_i)]\frac{1}{B_i} = \sum_{j=1}^{i} A_j$. The equation $\frac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial k_i} = \frac{\partial\gamma(\mathbf{q},\mathbf{k})}{\partial k_1}, \forall i \in \mathcal{N}_a$, means that the ISP has no incentives to adjust the capacity for each service class. By expanding this equation, we have

$$\sum_{j=1}^{i}[\rho(q_j) - \rho(q_{j-1})]\widetilde{R}'(\theta_j)\frac{\partial\Delta(q_i,k_i)/\rho(q_i)}{\partial k_i} = [\rho(q_1) - \rho(q_i)]\widetilde{R}'(\theta_1)\frac{\partial\Delta(q_1,k_1)/\rho(q_1)}{\partial k_1.}$$

Then, by rearranging terms, we have that $C_i\sum_{j=1}^{i} A_j = C_1A_1$. □

Proof of Corollary 3. By applying Theorem 3, we obtain the necessary optimality condition for q. Now we prove its uniqueness. Our objective to show that $\epsilon_{\bar{F}} = \epsilon_\rho/\epsilon_\Delta - 1$ has a unique solution q. It suffices to show that $\epsilon_{\bar{F}} - \epsilon_\rho/\epsilon_\Delta$ is increasing in q. Note that $\epsilon_{\bar{F}}$ is decreasing in p, and that $p = \bar{F}_V^{-1}(\Delta(q,1)/\rho(q))$ is decreasing in q. Hence, $\epsilon_{\bar{F}}$ is increasing in q. Note that ϵ_ρ is negative and non-increasing in q, and that ϵ_Δ is positive and non-increasing. Thus, $\epsilon_\rho/\epsilon_\Delta$ is non-increasing in q. Thus, we complete the proof. □

Proof of Theorem 4. We first prove the increasing property of the maximum revenue in capacity C. To facilitate the analysis, we first transform Problem 2. As we have shown in the proof of Lemma 2, $V_i = [p_{i+1}\rho(q_{i+1}) - p_i\rho(q_i)]/[\rho(q_{i+1}) - \rho(q_i)], \forall i$. By solving these equations, we have $p_i = \frac{1}{\rho(q_i)}\sum_{j=1}^{i}[\rho(q_j) - \rho(q_{j-1})]V_{j-1}$. By applying this result, we can transform Problem 2 into the following equivalent form:

$$\begin{aligned}
\underset{\mathbf{q},\mathbf{k},\mathbf{V}}{\text{maximize}} \quad & \sum_{i\in\mathcal{N}_a}[\rho(q_i) - \rho(q_{i-1})]V_{i-1}\bar{F}_V(V_{i-1}) \\
\text{subject to} \quad & \bar{F}_V(V_{i-1}) = \sum_{j=i}^{N}\Delta(q_j,k_j)/\rho(q_j), \qquad \forall i \in \mathcal{N}_a \\
& V_i > V_{i-1}, \qquad \forall i \in \mathcal{N}_a. \\
& \mathbf{q} \in \mathbb{R}_+^{N_a}, \mathbf{k} \in \mathcal{K}, \mathbf{V} \in \mathcal{V}^{N_a}.
\end{aligned}$$

Now the decision variables are the quality guarantee $\mathbf{q}$, the capacity $\mathbf{k}$, and the market segmentation $\mathbf{V} \triangleq (V_{i-1} : i \in \mathcal{N}_a)$. Given $\mathbf{q}, \mathbf{k}$, and $\mathbf{V}$, we next show how to increase the revenue if we increase the capacity by $\tilde{C} > 0$. We invest all capacity increased, i.e., $\tilde{C}$, to class N_a. Then, the capacity dividing vector becomes $\tilde{k}_i = \frac{k_iC}{C+\tilde{C}}, \forall i = 1,\ldots,N_a - 1$ and $\tilde{k}_{N_a} = \frac{k_{N_a}C+\tilde{C}}{C+\tilde{C}}$. As the capacity for class N_a increases, we can reduce q_{N_a} such that V_{N_a-1} remains unchanged, i.e., choosing $\tilde{q}_{N_a} < q_{N_a}$ such that $\Delta(\tilde{q}_{N_a},\tilde{k}_{N_a})/\rho(\tilde{q}_{N_a}) = \Delta(q_{N_a},k_{N_a})/\rho(q_{N_a})$. Then, it follows that if $(\mathbf{q},\mathbf{k},\mathbf{V})$ is an optimal solution before increasing the capacity, then the tuple $(\tilde{\mathbf{q}},\tilde{\mathbf{k}},\mathbf{V})$, where $\tilde{\mathbf{q}} = (q_1,\ldots,q_{N_a-1},\tilde{q}_{N_a})$ is an optimal solution after we increase the capacity. Then, the revenue improvement after increasing capacity is $(\rho(\tilde{q}_{N_a}) - \rho(q_{N_a}))V_{N_a-1}\bar{F}_V(V_{N_a-1}) > 0$. Hence, the maximum revenue increases with respect to the capacity.

Now we show that $\lim_{C\to\infty} R^*_{N_a} = \max_p p\bar{F}(p)$. First, observe that $V_i\bar{F}_V(V_i) \le \max_p p\bar{F}(p)$. Note that $\rho(q_i) > \rho(q_{i-1})$. Then, it follows that

$$\begin{aligned}
\sum_{i\in\mathcal{N}_a}[\rho(q_i) - \rho(q_{i-1})]V_{i-1}\bar{F}_V(V_{i-1}) &\le \sum_{i\in\mathcal{N}_a}[\rho(q_i) - \rho(q_{i-1})]\max_p p\bar{F}(p) \\
&= \rho(q_{N_a})\max_p p\bar{F}(p) < \max_p p\bar{F}(p)
\end{aligned}$$

This implies the following upper bound:

$$\lim_{C\to\infty} R^*_{N_a} \le \max_p p\bar{F}(p).$$

On the other hand, it is easy to observe that $\lim_{C\to\infty} R_1^* = \max_p p\bar{F}(p)$ and $R_{N_a}^* \geq R_1^*$. This implies the following lower bound:

$$\lim_{C\to\infty} R_{N_a}^* \geq \max_p p\bar{F}(p).$$

Hence, $\lim_{C\to\infty} R_{N_a}^* = \max_p p\bar{F}(p)$. □

Proof of Theorem 5. It is straightforward to see that increasing the maximum number of service classes N does not reduce the maximum revenue. Note that Theorem 4 does not depend on the number of active classes. Thus, the maximum revenue is $\max_p p\bar{F}(p)$. □

Proof of Theorem 6. The optimal solution $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ must satisfy the following KKT conditions:

$$\begin{cases} \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial p_i} + \displaystyle\sum_{j\in\mathcal{N}_a} u_j \frac{\partial D_j(\mathbf{p},\mathbf{q})}{\partial p_i} = 0, \forall i \in \mathcal{N}_a \\ \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial q_i} + \displaystyle\sum_{j\in\mathcal{N}_a} u_j \left[\frac{\partial D_j(\mathbf{p},\mathbf{q})}{\partial q_i} - \frac{\partial \Delta(q_j,k_j)}{\partial q_i}\right] = 0, \forall i \in \mathcal{N}_a \\ u_{N_a+1} \dfrac{\partial \sum_{j\in\mathcal{N}_a} k_j}{\partial k_i} - \displaystyle\sum_{j\in\mathcal{N}_a} u_j \frac{\partial \Delta(q_j,k_j)}{\partial k_i} = 0, \forall i \in \mathcal{N}_a \\ u_i \in \mathbb{R}, \forall i = 1,\ldots,N_a+1 \end{cases}$$

By eliminating the variables $u_1, \ldots, u_{N_a}$, we can simplify the above equations into the follow form:

$$\begin{cases} \dfrac{\partial R}{\partial p_i} + u_{N_a+1} \displaystyle\sum_{j\in\mathcal{N}_a} \frac{\partial D_j(\mathbf{p},\mathbf{q})}{\partial p_i} \Big/ \frac{\partial \Delta(q_j,k_j)}{\partial k_j} = 0, \forall i \in \mathcal{N}_a \\ \dfrac{\partial R}{\partial q_i} + u_{N_a+1} \displaystyle\sum_{j\in\mathcal{N}_a} \frac{\partial D_j(\mathbf{p},\mathbf{q})}{\partial q_i} \Big/ \frac{\partial \Delta(q_j,k_j)}{\partial k_j} - u_{N_a+1} \frac{\partial \Delta(q_j,k_j)}{\partial q_i} \Big/ \frac{\partial \Delta(q_j,k_j)}{\partial k_j} = 0, \forall i \in \mathcal{N}_a \\ u_{N_a+1} \in \mathbb{R} \end{cases}$$

Note that

$$\frac{\partial G}{\partial q_i} = \sum_{j\in\mathcal{N}_a} \frac{\partial D_j(\mathbf{p},\mathbf{q})}{\partial q_i} \Big/ \frac{\partial \Delta(q_j,k_j)}{\partial k_j} - D_i(\mathbf{p},\mathbf{q}) \frac{\partial^2 \Delta(q_i,k_i)}{\partial q_i \partial k_i} \Big/ \left(\frac{\partial \Delta(q_i,k_i)}{\partial k_i}\right)^2.$$

Then it follows that

$$\begin{cases} \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial p_i} + u_{N_a+1} \dfrac{\partial G}{\partial p_i} = 0, & \forall i \in \mathcal{N}_a \\ \dfrac{\partial R(\mathbf{p},\mathbf{q},\mathbf{k})}{\partial q_i} + u_{N_a+1} \left(\dfrac{\partial G}{\partial q_i} - H_i\right) = 0, & \forall i \in \mathcal{N}_a. \end{cases}$$

By eliminating the variable u_{N_a+1}, we reach the conclusion. □

Proofs of Theorems 7, 8 and 9. It is easy to verify that $(\xi\mathbf{p}, \mathbf{q}, \mathbf{k})$ is a feasible solution after scaling if and only if $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ is a feasible solution before scaling. Note that under solution $(\xi\mathbf{p}, \mathbf{q}, \mathbf{k})$ for the scaled problem, the market segment will also scale by ξ, as compared to the problem before scaling under solution $(\mathbf{p}, \mathbf{q}, \mathbf{k})$. Hence, the demand for each service class $D_i(\mathbf{p}, \mathbf{q})$ remains unchanged. We then obtain that the revenue will be scaled by ξ. This implies that if $(\xi\mathbf{p}, \mathbf{q}, \mathbf{k})$ optimal after scaling if and only if $(\mathbf{p}, \mathbf{q}, \mathbf{k})$ is optimal before scaling. Thus we reach the conclusion for Theorem 7. Similarly, we can prove Theorems 8 and 9. □

9 CONCLUSION

In this article, we propose an application-driven pricing scheme as a potential business model for the future Internet. It provides "pay as your service needs" flexibility so that content providers can flexibly choose particular service classes with various quality guarantees and prices for their applications. This pricing scheme can be beneficial to the entire Internet ecosystem since ISPs can increase revenues, CPs can deliver their contents with required quality guarantee, and end-users can improve their quality of experience. Our analytical and numerical investigations reveal that our scheme can significantly (by as high as over 20%) increase an ISP's revenue compared with the traditional single-class pricing, especially when: the ISP's capacity is limited; the applications are neutral to multiplexing; the reservation prices of applications are not very low; the number of CPs with strict quality requirement is moderately small.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful comments to improve the article.

REFERENCES

[1] Narayanaswamy Balakrishnan and Chin-Diew Lai. 2009. *Continuous bivariate distributions*. Springer Science & Business Media.
[2] Hemant K. Bhargava and Daewon Sun. 2008. Pricing under quality of service uncertainty: Market segmentation via statistical QoS guarantees. *European Journal of Operational Research* 191, 3 (2008), 1189–1204.
[3] Steven Blake, David Black, Mark Carlson, Elwyn Davies, Zheng Wang, and Walter Weiss. 1998. An architecture for differentiated services. *IETF RFC 2475* (1998).
[4] Timm Böttger, Félix Cuadrado, Gareth Tyson, Ignacio Castro, and Steve Uhlig. 2016. Open connect everywhere: A glimpse at the internet ecosystem through the lens of the netflix CDN. *CoRR* abs/1606.05519 (2016).
[5] Madeline Carr. 2016. Network neutrality. In *US Power and the Internet in International Relations*. 149–181.
[6] Chi-Kin Chau, Qian Wang, and Dah-Ming Chiu. 2014. Economic viability of paris metro pricing for digital services. *ACM Trans. Internet Technol.* 14, 2–3, Article 12 (Oct. 2014), 21 pages.
[7] Costas Courcoubetis, Laszlo Gyarmati, Nikolaos Laoutaris, Pablo Rodriguez, and Kostas Sdrolias. 2016. Negotiating premium peering prices: A quantitative model with applications. *ACM Trans. Internet Technol.* 16, 2, Article 14 (April 2016), 22 pages.
[8] Pascal Courty and Mario Pagliero. 2012. The impact of price discrimination on revenue: Evidence from the concert industry. *Review of Economics and Statistics* 94, 1 (2012), 359–369.
[9] Jon Crowcroft. 2007. Net neutrality: The technical side of the debate: A white paper. *SIGCOMM Comput. Commun. Rev.* 37, 1 (Jan. 2007), 49–56.
[10] R. Gibbens, R. Mason, and R. Steinberg. 2000. Internet service classes under competition. *IEEE Journal on Selected Areas in Communications* 18, 12 (Dec 2000), 2490–2498.
[11] P. Hande, M. Chiang, R. Calderbank, and S. Rangan. 2009. Network pricing and rate allocation with content provider participation. In *IEEE INFOCOM 2009*. 990–998.
[12] T. Henderson, J. Crowcroft, and S. Bhatti. 2001. Congestion pricing. Paying your way in communication networks. *IEEE Internet Computing* 5, 5 (2001), 85–89.
[13] Ravi Jain, Tracy Mullen, and Robert Hausman. 2001. Analysis of paris metro pricing strategy for QoS with a single service provider. In *Proceedings of the 9th International Workshop on Quality of Service (IWQoS'01)*. Springer-Verlag, London, UK, 44–58.
[14] R. T. B. Ma. 2016. Usage-based pricing and competition in congestible network service markets. *IEEE/ACM Transactions on Networking* 24, 5 (October 2016), 3084–3097.
[15] R. T. B. Ma and V. Misra. 2012. Congestion and Its role in network equilibrium. *IEEE Journal on Selected Areas in Communications* 30, 11 (December 2012), 2180–2189.
[16] J. K. MacKie-Mason and H. R. Varian. 1995. Pricing congestible network resources. *IEEE Journal on Selected Areas in Communications* 13, 7 (Sep 1995), 1141–1149.
[17] Peter Marbach. 1999. Pricing priority classes in a differentiated services network. In *Allerton Conference*, 25–9.
[18] P. Marbach. 2001. Pricing differentiated services networks: Bursty traffic. In *IEEE INFOCOM*, Vol. 2. 650–658.
[19] Eugenio Miravete. 2007. The limited gains from complex tariffs. CEPR Discussion Paper No. 4235.
[20] Vishal Misra. 2015. Routing money, not packets. *Commun. ACM* 58, 6 (May 2015), 24–27.

[21] Barrie R. Nault and Steffen Zimmermann. 2013. Policy, pricing and investment in a two-tier internet. In *Conference on Information Systems and Technology (CIST 2013)*.
[22] Andrew Odlyzko. 1999. Paris metro pricing for the internet. In *1st ACM Conference on Electronic Commerce (EC'99)*. ACM, New York, NY, USA, 140–147.
[23] I. C. Paschalidis and J. N. Tsitsiklis. 2000. Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking* 8, 2 (Apr. 2000), 171–184.
[24] Jon M. Peha. 2006. The benefits and risks of mandating network neutrality, and the quest for a balanced policy. In *34th Telecommunications Policy Research Conferences*.
[25] Vamseedhar Reddyvari Raja, Amogh Dhamdhere, Alessandra Scicchitano, Srinivas Shakkottai, Simon Leinen, et al. 2014. Volume-based transit pricing: Is 95 the right percentile?. In *International Conference on Passive and Active Network Measurement*. Springer, 77–87.
[26] David Reitman. 1991. Endogenous quality differentiation in congested markets. *The Journal of Industrial Economics* (1991), 621–647.
[27] David Ros and Bruno Tuffin. 2004. A mathematical model of the Paris Metro pricing scheme for charging packet networks. *Computer Networks* 46, 1 (2004), 73–85.
[28] S. Shakkottai, R. Srikant, A. Ozdaglar, and D. Acemoglu. 2008. The price of simplicity. *IEEE Journal on Selected Areas in Communications* 26, 7 (Sept. 2008), 1269–1276.
[29] J. Shu and P. Varaiya. 2003. Pricing network services. In *IEEE INFOCOM*, Vol. 2. 1221–1230.
[30] Xin Wang, Richard T. B. Ma, and Yinlong Xu. 2017. On optimal two-sided pricing of congested networks. *ACM Meas. Anal. Comput. Syst.* 1, 1, Article 7 (June 2017), 28 pages.
[31] Xin Wang and H. Schulzrinne. 2006. Pricing network resources for adaptive applications. *IEEE/ACM Transactions on Networking* 14, 3 (June 2006), 506–519.
[32] X. Wang, Y. Xu, and R. T. B. Ma. 2018. Paid peering, settlement-free peering, or both?. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*.
[33] Robert B. Wilson. 1993. *Nonlinear pricing*. Oxford University Press on Demand.
[34] Wired. 2016. How Amazon, Google, and Facebook Will Bring Down Telcos. https://www.wired.com/2016/12/the-end-of-telcos/.
[35] M. Zou, R. T. B. Ma, X. Wang, and Y. Xu. 2017. On optimal service differentiation in congested network markets. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 1–9.

Received March 2019; revised August 2019; accepted September 2019