

# Optimal Pricing, Scheduling, and Admission Control for Queueing Systems under Information Asymmetry

Tingting Cui, Ying-Ju Chen, and Zuo-Jun Max Shen\*  
University of California at Berkeley

September 11, 2009

## Abstract

In this paper, we attempt to provide a unified framework to study the optimal joint scheduling, admission control, and incentive compatible pricing mechanism. To this end, we study a problem setting in which a capacity-constrained service provider (server) modelled as an  $M/M/1$  queueing system intends to serve several segments of customers. Each customer requests the same amount of task, but they are heterogeneous in two attributes: their willingness to pay, and their willingness to wait, both of which are privately observed by this customer but are unknown to the server.

We show that a well-designed menu of probabilistic admission control along with priority pricing contracts may force customers to reveal their true valuations and at the same time induce customers that are more sensitive to the delay to opt for higher priorities. Thus, the probabilistic admission control allows the server to identify the customers that are willing to pay more for the service (thereby reducing the undesirable congestion), and consequently may enable the server to increase its revenue. Even though ex ante the server may exhibit specific preference over different groups of customers, the server may probabilistically admit more than one group to ensure incentive compatibility. Moreover, we find that randomized priority rule and strategic idleness can emerge as revenue maximizing solutions.

*Keywords:* congestion pricing, queueing systems, mechanism design

---

\*UC Berkeley, 4141 Etcheverry Hall, Berkeley, CA 94720; e-mail: {tingting, chen, shen}@ieor.berkeley.edu.

# 1 Introduction

Consider a capacity-constrained service provider (server) facing customers with different values for service and sensitivities to delays, both of which are their private information. What kind of pricing, scheduling, and admission control policy should the server follow in order to maximize her expected revenue? This question is faced by many firms in the production and service industries, including manufacturing, telecommunication and transportation. A common strategy adopted by these servers is to segment the customers by providing different classes of services. For example, many make-to-order manufacturers charge the customers based on the delivery dates, and transportation firms like Fedex and UPS offer a range of service classes from ground shipping to same day delivery. By offering the option to pay more for faster services, the server may extract more revenue from market segmentation. However, since the server only has aggregate information about the customer attributes but cannot tell apart individual customers, all customers can choose among all service classes in a self-interested way. This gives rise to the *incentive compatibility* issue, which the server must take into account in designing the revenue maximizing admission and scheduling policies.

This paper intends to provide a unified framework to study the aforementioned revenue maximization problem in the presence of asymmetric information regarding the customers' preferences. Our work is motivated by the recent papers of Afeche (2004), Yahalom et al. (2005), and Katta and Sethuraman (2005). Afeche (2004) adopts the mechanism design approach to evaluate the incentive compatible *priority pricing* problem in a queueing system where customers' valuations are drawn from a continuous distribution but their delay sensitivities can take only two values. He shows that the revenue-maximizing priority rule does not conform with the celebrated  $c\mu$  rule: it may require strategically inserted idleness, randomized priorities, or even reversed  $c\mu$  order. Yahalom et al. (2005) allow general distributions over valuation and convex delay cost; this implies that higher moments of delay may be influential in their context. Katta and Sethuraman (2005) impose perfect correlation between valuation and delay sensitivity. Thus, they are able to recast it as a (standard) single-dimensional adverse selection problem; consequently, they provide an efficient algorithm to characterize the optimal mechanism and find that pooling multiple types of customers into the same priority class emerges as an optimal solution.

Despite the insightful elaboration on the incentive issues and the managerial implications that

arise from those non-conventional queueing disciplines in the aforementioned work, an important feature of all these mechanisms is that *admission control is made through the design of priority pricing*. For example, in Katta and Sethuraman (2005) and Yahalom et al. (2005), only the priority classes (and the corresponding prices) are specified in the contracts. In Afeche (2004), the contract does specify the admission control. However, he focuses exclusively on the case with deterministic admission control. In other words, a customer is either admitted for sure or discarded entirely depending on the priority pricing scheme (the detailed discussions are deferred to Section 2). Thus, all the mechanisms in his model primarily use the priority classes as the sole screening tool to differentiate among customers.

In this paper, we argue that a previously ignored admission control policy plays a significant role in mitigating the information asymmetry between the server and the customers. Specifically, we show that a well-designed menu of admission control along with priority pricing contracts may force customers to reveal their true valuations; at the same time, this menu also induces the customers that are more sensitive to the delay to opt for higher priorities. The intuition is as follows. The customers with high valuations have higher opportunity costs when they do not get the services. Hence, if a *probabilistic* admission control policy is used (with different probability of rejecting customers), customers with high valuations may be willing to pay more for a better chance of getting admitted. Thus, the probabilistic admission control allows the server to choose the right customers to serve (thereby reducing the undesirable congestion) and consequently may enable the server to receive more revenue from those customers.

We illustrate our idea in a stylized model in which both the valuation and the delay sensitivity can take only two values – high or low. This allows us to classify the customers into four groups (types):  $\{LH, HH, LL, HL\}$ , where the first component specifies whether the valuation is high ( $H$ ) or low ( $L$ ), and the second component depicts whether the customer is highly sensitive to the delay ( $H$  in this case) or not ( $L$ ). While not attempting to be all inclusive and the most general possible, this four-type model allows us to derive concrete managerial implications. Specifically, we show that the server may partially admit (through probabilistic admission) more than one customer types, although ex ante one type is more favorable than the other. Moreover, the server may assign different/randomized priorities for customers with same delay sensitivity but different valuations for service. Finally, the optimal contracts may require strategically inserted idleness to ensure incentive compatibility, which echoes the results of Afeche (2004).

Admission control has long been recognized as a valuable tool to balance the throughput and congestion trade-off in queueing systems; see Stidham (1985) and Stidham (2002) for good surveys. The probabilistic admission policy we propose is widely adopted by connection admission control (CAC) protocols in communication networks; see, e.g., Gibbens et al. (1995) and Lewis et al. (1998) and the references therein. An example is the RSVP (Resource Reservation Protocol) used for the ATM (Asynchronous Transfer Mode) network. In this protocol, different groups of consumers (packets such as email, ftp, voice data, etc.) are given choices over a number of flags (classes); each class is associated with a price, a priority class, and the probability of being dropped (that is analogous to the probabilistic admission control). See, e.g., Chang and Petr (2001) and Zhang et al. (1993) for the detailed descriptions of the protocol. (These protocols are proposed primarily from the system efficiency standpoint. On the contrary, in our model, the joint admission control/priority pricing is adopted to maximizing the server’s revenue.)

Our model falls in the category of mechanism design problems with *multi-dimensional* private information (willingness to pay and willingness to wait) and screening tools (admission control and priority classes). Multi-dimensional mechanism design problems have long been recognized to be notoriously complicated and sometimes analytically intractable. The main challenges arise from the lack of complete ordering among the multi-dimensional types. Unlike the classical uni-dimensional framework, there is simply no unified way to *ex ante* identify redundant/ binding incentive compatibility constraints, thereby breaking down the systematic approach that has been prominently adopted in the literature; see the recent survey by Rochet and Stole (2005). Moreover, the unique capacity constraint that arises from our queueing framework brings in new challenges and results in novel insights that would not occur in other contexts.

Our paper is related to the vast literature on pricing, scheduling, and admission control in queueing systems. Classical papers in this field typically treat this problem in a centralized manner (i.e., a central planner is able to control all the behavior of the server, customers, etc.); see, e.g., Coffman and Mitrani (1980), Shanthikumar and Yao (1992), and Stidham (2002) for an excellent survey. In contrast, we incorporate the strategic customer behavior and asymmetric information. The strategic customer behavior has also been incorporated in the design of queueing systems at least dating back to Naor (1969); see the monograph by Hassin and Haviv (2002) for a review of this literature. Mendelson (1985) and Mendelson and Whang (1990) are among the first to study socially optimal and incentive compatible priority pricing strategies in queueing systems. As

aforementioned, Afeche (2004), Katta and Sethuraman (2005), and Yahalom et al. (2005) focus on incentive compatible priority pricing policies that maximizes the server’s revenue. In line with this research stream, we introduce the freedom of choosing the (probabilistic) admission control that allows the server to extract more revenue from the customers effectively. Furthermore, by incorporating the possibility of probabilistic admission control, this paper expands the multi-dimensional nature of the classical incentive compatible revenue management to its full force.

Since we adopt the mechanism design approach to study this joint pricing, scheduling, and admission control problem, our work is also related to the principal-agent problems in which the principal intends to design an appropriate mechanism (contract) for the agents with private information to self-select. This framework has been extensively applied to various contexts in the operations research field, mostly within the single-dimensional framework, including capacity allocation (Cachon and Lariviere (1999)), supplier-retailer contracting (Corbett and de Groote (2000), Ha (2001)), product specification and production planning (Iyer et al. (2005)), inventory risk mitigation through promised lead time (Lutze and Ozer (2008)), pricing information goods (Wu and Chen (2008)), long-term contract design (Zhang and Zenios (2008)), and supply chain disruptions (Yang et al. (2009)). In contrast with the aforementioned papers, the *multi-dimensional* nature of our queueing framework inevitably creates new challenges. Wilson (1993) and Armstrong (1996) are the first to solve the multi-dimensional problems in closed form under specific assumptions on the model characteristics. Armstrong and Rochet (1999) provide a unified algorithm to solve discrete (specifically, four-type) multi-dimensional problems. This four-type framework is later adopted by Armstrong (2000) and Asker and Cantillon (2009) to study the forward and reverse auctions. Our four-type framework is also motivated by this stream of research. Nevertheless, the unique resource constraint that arises from the queueing framework results in a number of novel insights/ phenomena that would not occur in other contexts.

The remainder of this paper is organized as follows. In Section 2, we describe the model setup. In Section 3, we present socially optimal contracts under symmetric information, as a benchmark for our study of information asymmetry. We discuss revenue maximizing contracts under asymmetric information in Section 4, with structural properties of the optimal solutions in Section 4.1, full characterization of the exact optimal mechanism for some special cases in Section 4.2, and novel features of the revenue maximizing contracts in Section 4.3. In Section 5, we demonstrate the revenue gains from the admission control policy using two numerical examples. We summarize our

findings and give future research directions in Section 6. We include all the proofs of propositions and lemmas in the appendix, and relegate the detailed derivations for the special cases to the online appendix.

## 2 Formulation

We consider a stylized model in which a capacity-constrained server modelled as an  $M/M/1$  queueing system intends to serve several segments of customers. Customers request the same amount of task but are heterogeneous in two attributes: their willingness to pay, and their willingness to wait. Specifically, we assume that the service time of each customer follows an exponential distribution with a common rate  $\mu$ . Nevertheless, their valuations, denoted by  $v$ , and their delay sensitivities, denoted by  $c$ , are different across different groups. The value attribute  $v > 0$  characterizes the customer's willingness to pay (in the absence of delay) for one unit of service, while the delay sensitivity  $c > 0$  specifies the penalty per unit of time while the customer is kept in the system (service time included).

We assume that both attributes can take only two values to simplify our analysis. Specifically, we assume that  $v \in \{v_L, v_H\}$  and  $c \in \{c_L, c_H\}$ , where  $\Delta_v = v_H - v_L > 0$  and  $\Delta_c = c_H - c_L > 0$ . Given these values, there are four combinations  $\{(v_L, c_H), (v_H, c_H), (v_L, c_L), (v_H, c_L)\}$ , which are denoted by  $LH$ ,  $HH$ ,  $LL$ , and  $HL$ , respectively in the sequel. A customer with valuation  $v_H$  is willing to pay more for the service than the one with  $v_L$ ; likewise, a customer endowed with a delay sensitivity  $c_H$  incurs a higher penalty (compared to the case with  $c_L$ ). Each group of customers arrive at the system following a Poisson process, and we use  $\lambda_{ij}$  to denote the aggregate arrival rate of group  $ij$  of customers, where  $ij \in \{LH, HH, LL, HL\} \equiv T$ . Notably, from the server's viewpoint, type- $HL$  customers are the most favorable customers since they are willing to pay more for the service and do not mind waiting so much. On the contrary, the server can extract the least amount of profit from the type- $LH$  customers due to their low willingness to pay and high delay sensitivity. In compliance with the literature on incentive compatible priority pricing, we assume that the arrival process, the value and cost distributions and the service procedure are common knowledge. However, a customer's valuation and delay sensitivity are privately observed by this customer but unknown to the server. Thus, this private preference profile also represents a customer's *type*.

The server's problem is to design an appropriate mechanism to maximize his long-run expected payoff. In the absence of the information about the customers' preference, the server faces an adverse selection problem. As suggested by the agency literature (Laffont and Martimort (2002)), a common approach is to offer the customers a *menu* of contracts and let her self-select. Furthermore, the revelation principle allows us to restrict our attention to the direct mechanism in which the server simply requests the customers to report their types and then choose the contracts on behalfs of the customers. Thus, we assume that the server offers a menu of contracts  $\{q_{ij}, w_{ij}, p_{ij}\}$ , where  $q_{ij} \in [0, 1]$  is the admission rate,  $w_{ij}$  is the expected delay, and  $p_{ij}$  denotes the associated price charged by the server. Upon arrival, each customer decides which service class to purchase, and is charged and scheduled as prescribed by the contract. Given a contract  $\{q, w, p\}$ , a type- $ij$  customer receives an expected (net) utility:  $q(v_i - c_j w - p)$ . We assume that a customer receives a null (zero) expected utility upon walking away without loss of generality. Notably, based on the above descriptions, the server is allowed to adopt a *stochastic/probabilistic* admission control policy for a specific type (this occurs when  $q_{ij} \neq \{0, 1\}$ ). This is in strict contrast with the extant literature on priority pricing, see, e.g., Afeche (2004), Katta and Sethuraman (2005), and Yahalom et al. (2005).

We restrict our attention to static scheduling policy and allow preemption. Furthermore, we adopt the achievable-region approach introduced by Coffman and Mitrani (1980) and Shanthikumar and Yao (1992). Under our queueing framework, the associated expected delay is confined with the following resource constraints:

$$\sum_{ij \in T} \lambda_{ij} q_{ij} < \mu, \text{ and } \sum_{ij \in S} \frac{\lambda_{ij} q_{ij} w_{ij}}{\mu} \geq \frac{\sum_{ij \in S} \lambda_{ij} q_{ij} / \mu}{\mu - \sum_{ij \in S} \lambda_{ij} q_{ij}}, \forall S \subseteq T. \quad (\text{RE})$$

The condition  $\sum_{ij \in T} \lambda_{ij} q_{ij} < \mu$  guarantees that the system size does not explode (since the effective aggregate arrival rate  $\sum_{ij \in T} \lambda_{ij} q_{ij}$  is less than the service rate). In the second inequality of (RE), the left-hand side is the expected steady-state virtual load (defined as the remaining processing time) in set  $S$  as we recall by Little's law that  $\lambda_{ij} w_{ij}$  is the steady-state queue length of group  $ij$ ; the right-hand side corresponds to the average sojourn time (the waiting time plus the service time) when the customers in set  $S$  are given the *absolute* priority over all other customers outside this set (i.e.,  $T \setminus S$ ). Thus, from the viewpoint of the customers in set  $S$ , it is as if those customers outside this set are never in the system (and thus no further congestion is incurred due to the presence of customers in  $T \setminus S$ ). This derivation follows from the classical queueing theory (see, e.g., Coffman and Mitrani (1980), Shanthikumar and Yao (1992) and also Katta and Sethuraman (2005)). For ease of notation, we denote  $RE(S)$  as the resource constraint associated with the set  $S$ .

It is worth mentioning that this achievable region can be regarded as a sort of resource constraints for this queueing system. Moreover, each extreme point of this achievable region corresponds to a specific absolute priority rule. In our four-group setting, an extreme point is determined by four binding constraints, each of which is associated a specific set. Further, these sets must be *nested* in order to avoid any conflict between the queueing discipline. For example, if the sets associated with an extreme point is  $\{HH\}$ ,  $\{HH, LH\}$ ,  $\{HH, LH, HL\}$ , and  $\{HH, LH, HL, LL\}$ , the corresponding absolute priority rule is  $HH, LH, HL, LL$ , in descending order. This peculiar property implies that the achievable region is a base of a *polymatroid* (Shanthikumar and Yao (1992)). Another interesting observation is that since any interior point can be represented as a convex combination of a finite number of extreme points and feasible directions (which correspond to the “*strategic idleness*” in the terminology of Afeche (2004)). This convex combination also gives rise to a detailed implementation through a (*randomized*) priority rule, i.e., a certain group of customers are given priority only probabilistically. See Shanthikumar and Yao (1992) for more discussions and algorithms that implement the priority rules.

By the revelation principle, we restrict our attention to direct revelation mechanisms. Let

$$u(i'j'|ij) = q_{i'j'}(v_i - c_j w_{i'j'} - p_{i'j'})$$

denote the expected utility of a type- $ij$  customer who pretends to be type- $i'j'$ . For ease of notation, define

$$W_{ij} = q_{ij}w_{ij}, \text{ and } P_{ij} = q_{ij}p_{ij},$$

the customer's expected utility can be rewritten as  $u(i'j'|ij) = v_i q_{i'j'} - c_j W_{i'j'} - P_{i'j'}$ . In order to induce customers to participate, the following *individual rational* (IR) constraint has to hold:

$$v_i q_{ij} - c_j W_{ij} - P_{ij} \geq 0, \forall ij \in T, \tag{IR}$$

where the right-hand side corresponds to the customers' reservation utility (which is normalized to zero). Furthermore, the menu of contracts has to induce the customers to willingly reveal their types, thereby giving rise to the following *incentive compatible* (IC) constraint:

$$v_i q_{ij} - c_j W_{ij} - P_{ij} \geq v_i q_{i'j'} - c_j W_{i'j'} - P_{i'j'}, \forall ij, i'j' \in T, \tag{IC}$$

where the left-hand side, as aforementioned, is the expected utility of a type- $ij$  customer under truth-telling, and the left-hand side corresponds to the case of misrepresentation. Note that even if a customer misreports her type, the actual valuation as well as the delay sensitivity remain genuine

( $v_i$  and  $c_j$ , respectively). We use  $IC(ij - i'j')$  to denote the incentive compatibility constraint that guarantees that a type- $ij$  customer does not want to pretend to be type- $i'j'$ .

Having discussed the customers' incentive problems, we now turn to the server's side. The server's goal is to find an appropriate menu of contracts that maximize her expected revenue:

$$\begin{aligned} \max_{\{q_{ij}, w_{ij}, p_{ij}\}} \quad & \sum_{ij \in T} \lambda_{ij} P_{ij}, \\ \text{s.t.} \quad & \text{(IC), (IR), and (RE)}. \end{aligned}$$

For our convenience, we can replace the decision variables  $\{q_{ij}, w_{ij}, p_{ij}\}$  by  $\{q_{ij}, W_{ij}, P_{ij}\}$  following the definitions of  $W_{ij}$  and  $P_{ij}$ . Moreover, we introduce the “*information rent*”:

$$R_{ij} \equiv v_i q_{ij} - c_j W_{ij} - P_{ij}$$

for each  $ij \in T$ . From the definition of  $R_{ij}$ , we have  $P_{ij} = v_i q_{ij} - c_j W_{ij} - R_{ij}$ . After these substitutions, the server's problem is reformulated as below:

$$\begin{aligned} \max_{\{q_{ij}, W_{ij}, R_{ij}\}} \quad & \sum_{ij \in T} \lambda_{ij} (v_i q_{ij} - c_j W_{ij} - R_{ij}) \\ \text{s.t.} \quad & R_{ij} - R_{i'j'} \geq (v_i - v_{i'}) q_{i'j'} - (c_j - c_{j'}) W_{i'j'}, \quad \forall ij, i'j' \in T, \end{aligned} \quad (1)$$

$$R_{ij} \geq 0, \quad \forall ij \in T, \quad (2)$$

$$\sum_{ij \in T} \lambda_{ij} q_{ij} < \mu, \quad (3)$$

$$\sum_{ij \in S} \lambda_{ij} W_{ij} \geq \frac{\sum_{ij \in S} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in S} \lambda_{ij} q_{ij}}, \quad \forall S \subseteq T, \quad (4)$$

$$W_{ij} \geq 0, \quad 0 \leq q_{ij} \leq 1, \quad \forall ij \in T,$$

where (1) follows from (IC), (2) follows from (IR), and (3) and (4) are simply a restatement of (RE).

In the sequel, we derive the optimal mechanism (from the server's perspective) with four groups of customers. As a benchmark, we first study optimal (socially optimal) contracts under symmetric information in Section 3. Revenue maximizing contracts under asymmetric information are discussed in Section 4.

### 3 Optimal Contracts under Symmetric Information

To demonstrate the impact of information asymmetry, we first derive the optimal menu of contracts when the server has perfect knowledge of the customers' type information. We refer to this benchmark case as the scenario with symmetric information.

As the server knows the customers' types, the incentive compatibility condition (1) is no longer required, and since the server can extract the entire social surplus, the problem reduces to social maximization. The optimal contract design problem in this scenario can be formulated as follows:

$$\max_{\{q_{ij}, W_{ij}\}} \sum_{ij \in T} \lambda_{ij} (v_i q_{ij} - c_j W_{ij}) \quad (5)$$

$$\text{s.t.} \quad \sum_{ij \in T} \lambda_{ij} q_{ij} < \mu, \quad (6)$$

$$\sum_{ij \in S} \lambda_{ij} W_{ij} \geq \frac{\sum_{ij \in S} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in S} \lambda_{ij} q_{ij}}, \quad \forall S \subseteq T, \quad (7)$$

$$W_{ij} \geq 0, \quad 0 \leq q_{ij} \leq 1, \quad \forall ij \in T. \quad (8)$$

Let  $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$  be an optimal solution to (5)-(8). The following propositions characterize the optimal admission control and priority ranking policies under symmetric information.

#### 3.1 Admission Preference

As a profit maximizer, the server has preferences of admitting certain types of customers over the others. We say that the server has *strong preference* of type  $ij$  over type  $i'j'$  if she does not admit any type  $i'j'$  customers unless she fully admits all type  $ij$  customers; i.e.  $q_{i'j'} = 0$  if  $q_{ij} < 1$ . The following proposition characterizes the component-wise strong preference of a socially optimal admission policy: among customers with the same delay sensitivity, the server has strong preference of types with higher valuation; analogously, among customers with same valuation, the server has strong preference of types with lower delay sensitivity.

**Proposition 1.** *A socially optimal menu of contracts  $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$  satisfies the following properties:*

$$\hat{q}_{LL} = 0, \text{ if } \hat{q}_{HL} < 1;$$

$$\hat{q}_{LH} = 0, \text{ if } \hat{q}_{HH} < 1;$$

$$\hat{q}_{HH} = 0, \text{ if } \hat{q}_{HL} < 1;$$

$$\hat{q}_{LH} = 0, \text{ if } \hat{q}_{LL} < 1.$$

For comparison, we define the *weak preference* of customer types to be the ordering of their admission probabilities; i.e. the server has weak preference of type  $ij$  customers over type  $i'j'$  customers if and only if  $q_{ij} > q_{i'j'}$ . It is easy to see that the strong preference always implies the weak preference. Later on, we will show that component-wise strong preference no longer holds in revenue maximizing contracts under information asymmetry; however, weak preference is preserved.

### 3.2 Priority Scheduling

We say that type  $ij$  customers has *absolute priority* over type  $i'j'$  customers if type  $ij$  always has the preemptive advantage of service over type  $i'j'$ . In terms of the resource constraints, there exists  $S \subseteq T$  such that  $ij \in S$ ,  $i'j' \notin S$ , and  $RE(S)$  is binding. We say that type  $ij$  has *randomized priority* over type  $i'j'$  if type  $ij$  customers have shorter lead times than type  $i'j'$ , but does not have absolute priority over type  $i'j'$ . Finally, we say that type  $ij$  and type  $i'j'$  have *equal priority* if their average lead times are the same, i.e.,  $w_{ij} = w_{i'j'}$ .

The following proposition states that in a socially optimal contract, customers with higher delay sensitivity have absolute priority over the others; however, there's no need to differentiate customers with same delay sensitivity but different valuations. On the contrary, while the first assertion still holds in a revenue maximizing contract under information asymmetry, it may be optimal to assign absolute or randomized priorities among customers with same delay sensitivity but different valuation.

**Proposition 2.** *There exists a socially optimal menu of contracts  $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$  that satisfies the following*

properties:

$$\frac{\hat{W}_{LH}}{\hat{q}_{LH}} = \frac{\hat{W}_{HH}}{\hat{q}_{HH}} < \frac{\hat{W}_{LL}}{\hat{q}_{LL}} = \frac{\hat{W}_{HL}}{\hat{q}_{HL}};$$

$$\lambda_{LH}\hat{W}_{LH} + \lambda_{HH}\hat{W}_{HH} = \frac{\lambda_{LH}\hat{q}_{LH} + \lambda_{HH}\hat{q}_{HH}}{\mu - \lambda_{LH}\hat{q}_{LH} - \lambda_{HH}\hat{q}_{HH}}.$$

### 3.3 Work Conservation

We say that a scheduling policy follows the *work conservation* rule if it never idles the server. In terms of the resource constraints,  $RE(T)$  is always binding in any work conservation policy. Since we allow preemption, a socially optimal contract that minimizes delay costs should always satisfy the work conservation condition. However, in the presence of information asymmetry, it may be optimal to insert unforced idleness to delay the service for customers with lower delay sensitivity, in order to induce customers with higher delay sensitivity to report their true types.

**Proposition 3.** *A socially optimal menu of contracts  $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$  satisfies*

$$\sum_{ij \in T} \lambda_{ij} \hat{W}_{ij} = \frac{\sum_{ij \in T} \lambda_{ij} \hat{q}_{ij}}{\mu - \sum_{ij \in \{LH, HH\}} \lambda_{ij} \hat{q}_{ij}}.$$

The proof of Propositions 1-3 is straightforward and thus is omitted for conciseness. In the next Section, we present solutions for the optimal contract design problem under asymmetric information, focusing on the novel features introduced by information asymmetry.

## 4 Optimal Contracts under Information Asymmetry

In this section, we first overview general properties of the server's optimal contracts under information asymmetry in Section 4.1. Then we discuss detailed solutions for different cases of parameter values in Section 4.2. In Section 4.3, we compare the optimal contracts to that under symmetric information.

### 4.1 General Properties of the Optimal Contracts

To characterize structural properties of the revenue maximizing contracts, we use  $q^* \in \mathcal{R}_+^{|T|}$  to denote the optimal allocation rule and  $W^* \in \mathcal{R}_+^{|T|}$  to denote the corresponding optimal schedule. We first focus on the admission control policies and summarize our results in the next two

propositions. Proposition 4 shows that the server has strong preference of customers with higher evaluation among all that with lower delay sensitivity, and customers with lower delay sensitivity among all that with higher evaluation (*strong preference at top*).

**Proposition 4.** *A revenue maximizing menu of contracts  $(\mathbf{q}^*, \mathbf{W}^*, \mathbf{R}^*)$  has the following properties:*

$$\begin{aligned} q_{LL}^* &= 0, \text{ if } q_{HL}^* < 1; \\ q_{HH}^* &= 0, \text{ if } q_{HL}^* < 1. \end{aligned}$$

However, as demonstrated by the special cases in Section 4.2, strong preference no longer holds among customers with lower evaluations or higher delay sensitivities, as the server may partially admit both the  $HH$  and  $LH$  types or the  $LL$  and  $LH$  types. Nonetheless, Proposition 5 shows the monotonicity on the optimal admission probability among these types (*weak preference at bottom*).

**Proposition 5.** *The following assertions hold for any revenue maximizing menu of contracts  $(\mathbf{q}^*, \mathbf{W}^*, \mathbf{R}^*)$ :*

$$\begin{aligned} q_{LH}^* &\leq q_{HH}^*; \\ q_{LH}^* &\leq q_{LL}^*. \end{aligned}$$

Next, we turn to the priority scheduling policy. Analogously to the case under symmetric information, Proposition 6 shows that an optimal menu of contracts should always grant customers with higher delay sensitivity the absolute service priority.

**Proposition 6.** *A revenue maximizing solution  $(\mathbf{q}^*, \mathbf{W}^*, \mathbf{R}^*)$  satisfies*

$$\lambda_{HH}W_{HH}^* + \lambda_{LH}W_{LH}^* = \frac{\lambda_{HH}q_{HH}^* + \lambda_{LH}q_{LH}^*}{\mu - \lambda_{HH}q_{HH}^* - \lambda_{LH}q_{LH}^*}.$$

As opposed to the symmetric information case, the server may have to use more than two priority classes in order to differentiate customers with the same delay sensitivity but different valuations. Proposition 7 shows that the  $HH$  type should have absolute priority over the  $LH$  type if neither of them are fully admitted.

**Proposition 7.** *The resource constraint  $RE(\{HH\})$  is binding if  $q_{HH}^* < 1$ ; i.e.,*

$$\lambda_{HH}W_{HH}^* = \frac{\lambda_{HH}q_{HH}^*}{\mu - \lambda_{HH}q_{HH}^*}.$$

Depending on the system configurations, it may be optimal to use randomized or reversed priority ranking between the  $HH$  and the  $LH$  types. It is also possible that the optimal contract schedules the  $HL$  and  $LL$  types at different priorities, along with strategic idleness to ensure incentive compatibility. These interesting phenomena are demonstrated by the special cases discussed in Section 4.2, and are discussed in detail in Section 4.3.

## 4.2 Some special cases

In this section, we present some special cases for which intriguing phenomena arise in terms of the optimal admission control and priority rules. Following Proposition 4 and 5, the server has strong preference of the  $HL$  type over any other customer types. In other words, the server will not admit customers of any other types if he does not fully admit the  $HL$  type. However, if the server fully admits the  $HL$  type, in general he could partially admit all three inferior types of customers, since the server has no clear-cut preference between the customers of two (intermediate) types  $HH$  and  $LL$ , although type- $HL$  (type- $LH$ ) customers are always the most (least) favorable from the seller's perspective. For simplicity, we focus on the extreme cases in which the server partially admits at most two customer types.

When  $\frac{\Delta_c}{\Delta_v}$ , the ratio of the difference of delay sensitivity and the valuation difference, exceeds certain threshold level, upon fully admitting the most favorable type- $HL$  customers, the server admits the type- $LL$  customers until they are exhausted before admitting the type- $HH$  and  $LH$  customers. When  $\frac{\Delta_c}{\Delta_v}$  is below certain critical level, the preference is reversed: the server now intends to exhaust the type- $HH$  customers before admitting any type- $LL$  and  $LH$  customers. It would be difficult to determine the ranking between these two types in the intermediate cases; thus, in the sequel, we restrict ourselves to these two extreme cases.

Our results are summarized in Table 1, which are categorized by the the number of admitted groups (types) in the optimal solution. For each special case, we specify the admission policy, the number of priority classes required, and the detailed scheduling policy in each case. In the sequel we classify them by the number of admitted groups and elaborate on these cases. Detailed derivations of the optimal solutions for the special cases are provided in the online appendix.

Table 1: Summary of optimal contracts under asymmetric information

Number of admitted classes	$\frac{\Delta_c}{\Delta_v} \geq \mu(1 + \frac{\lambda_{HL}}{\lambda_{LL}})$	$\frac{\Delta_c}{\Delta_v} \leq \frac{(\mu - \lambda_{HH})^2}{\mu(1 + \lambda_{HL}/\lambda_{LL})}$
1	<p><b>Case 1a</b></p> <p><math>q_{HL} = 1, q_{LL} = q_{HH} = q_{LH} = 0</math></p> <p>Single queue.</p>	<p><b>Case 1b</b></p> <p><math>q_{HL} = 1, q_{HH} = q_{LL} = q_{LH} = 0</math></p> <p>Single queue.</p>
2	<p><b>Case 2a</b></p> <p><math>q_{HL} = 1, 0 &lt; q_{LL} \leq 1, q_{HH} = q_{LH} = 0</math></p> <p>Single queue, <math>w_{HL} = w_{LL}</math>.</p>	<p><b>Case 2b</b></p> <p><math>q_{HL} = 1, 0 &lt; q_{HH} \leq 1, q_{LL} = q_{LH} = 0</math></p> <p>Two queues, <math>w_{HH} &lt; w_{HL}</math>.</p> <p>Absolute priority rule.</p>
3	<p><b>Case 3a</b></p> <p><math>q_{HL} = q_{LL} = 1, 0 &lt; q_{HH} \leq 1, q_{LH} = 0</math></p> <p>Two queues, <math>w_{HH} &lt; w_{HL} = w_{LL}</math>.</p> <p>Absolute priority rule.</p>	<p><b>Case 3b</b></p> <p><math>q_{HL} = q_{HH} = 1, 0 &lt; q_{LL} \leq 1, q_{LH} = 0</math></p> <p>Three queues, <math>w_{HH} &lt; w_{HL} &lt; w_{LL}</math>.</p> <p>Randomized priorities between <i>LL</i> and <i>HL</i>.</p> <p>Strategic idleness may be optimal.</p>
4	<p><math>q_{HL} = q_{LL} = 1, 0 &lt; q_{LH} \leq q_{HH} \leq 1</math></p> <p><b>Case 4a-1</b></p> <p>If <math>q_{HH} &lt; 1</math>, three queues.</p> <p><math>w_{HH} &lt; w_{LH} &lt; w_{HL} = w_{LL}</math>,</p> <p>Absolute priority rule.</p> <p><b>Case 4a-2</b></p> <p>If <math>q_{HH} = 1</math> and <math>q_{LH} &lt; 1</math>, three queues.</p> <p><math>w_{HH} &lt; w_{LH} &lt; w_{HL} = w_{LL}</math>.</p> <p>Randomized priorities between <i>HH</i> and <i>LH</i>.</p> <p><b>Case 4a-3</b></p> <p>If <math>q_{HH} = q_{LH} = 1</math>, two queues.</p> <p><math>w_{HH} = w_{LH} &lt; w_{HL} = w_{LL}</math>.</p> <p>Absolute priority rule.</p>	<p><math>q_{HL} = q_{HH} = 1, 0 &lt; q_{LH} \leq q_{LL} \leq 1</math></p> <p><b>Case 4b-1</b></p> <p>If <math>q_{LL} &lt; 1</math>, four queues.</p> <p><math>w_{LH} &lt; w_{HH} &lt; w_{HL} &lt; w_{LL}</math></p> <p>Randomized priorities between <i>LL</i> and <i>HL</i>.</p> <p>Strategic idleness may be optimal.</p> <p><b>Case 4b-2</b></p> <p>If <math>q_{LL} = 1</math> and <math>q_{LH} &lt; 1</math>, three queues.</p> <p><math>w_{LH} &lt; w_{HH} &lt; w_{HL} = w_{LL}</math>.</p> <p>Randomized priorities between <i>LH</i> and <i>HH</i>.</p> <p>Strategic idleness may be optimal.</p> <p><b>Case 4b-3</b></p> <p>If <math>q_{LL} = q_{LH} = 1</math>, two queues.</p> <p><math>w_{LH} = w_{HH} &lt; w_{HL} = w_{LL}</math>.</p> <p>Absolute priority rule.</p>

### 4.2.1 One group of customers

Let us start with the simplest case in which only one group of customers are admitted (Case 1a and 1b). Since the server can extract the most revenue from the type-*HL* customers (who values the service highly and is less averse to the delay), this is the only group of customers that are admitted (as shown in Proposition 4). In this case, the “priority rule” degenerates since all admitted customers are identical.

### 4.2.2 Two groups of customers

A slightly more interesting case is when the server admits two groups of customers. By the above arguments, in addition to the type-*HL* customers, when the valuation difference is relatively small, the other admitted type is the type-*LL* (Case 2a); whereas the type-*HH* customers are admitted if the difference of delay sensitivity is relatively small (Case 2b).

In Case 2a, the two admitted types – *HL* and *LL* – have the same delay sensitivity. Thus, it makes no sense for the server to provide different priority rules; the only relevant parameter to differentiate between these two types is the admission probability. On the contrary, the server admits two types with different delay sensitivities (*HL* and *HH*) in Case 2b. In such a scenario, offering two priority classes allows the server to differentiate between them, since the type-*HH* customers are more averse to the delay and therefore are willing to pay more for a higher priority. We therefore observe that the server offers two priority classes in this case.

### 4.2.3 Three groups of customers

When the server admits three groups of customers, the admission and scheduling rules again critically depend on  $\frac{\Delta_c}{\Delta_v}$ ; nevertheless, the set of admitted groups (*HL*, *HH*, and *LL*) is the same in the two extreme cases. When the valuation difference is relatively small (Case 3a), the server fully admits type *LL* and probabilistically admits type *HH* customers; the preference is reversed and the *LL* type is partially admitted if the difference of delay sensitivity is relatively small (Case 3b).

Following from Proposition 6, type-*HH* customers are given absolute priority over other types in both cases. However, the priority ranking between the *HL* and the *LL* types differs in the two extreme cases. In Case 3a, there is no need to offer different priority classes for the *HL* and

$LL$  types, since doing so will affect neither the system delay cost nor the customers' incentives. However, in Case 3b, the server intends to assign the type- $HL$  customers a higher priority than the type- $LL$  customers, even if these customers are homogeneous in terms of their delay sensitivity. In such a scenario, the purpose of this delay differentiation is to prevent the type- $HH$  customers from misrepresenting themselves as either the  $LL$  or the  $HL$  type. The randomized priority rule provides the server with the desired flexibility to align the incentives of the customer's and minimize the information rent.

We further find that occasionally the server may insert unforced idleness to the queues with lower priority; in other words, *strategic idleness* may emerge as an optimal solution. This is because all work-conserving priority rules, although achieving system efficiency, result in severe incentive compatibility issues. Consequently, in order to differentiate the customers, the server must reduce the prices for the high priority queue significantly. Alternatively, the server may be better off by distorting the queueing discipline rather than adjusting the prices dramatically. This observation echoes the seminal work by Afeche (2004).

#### 4.2.4 Four groups of customers

Finally, let us consider the case in which all four groups of customers are admitted, when the difference in delay sensitivity is relatively more significant (Case 4a-1 to 4a-3). We find that given the (soft) resource constraint, it may be in the server's best interest to probabilistically admit both the  $HH$  and  $LH$  types, even if ex ante the type- $LH$  customers are perceived as the worst group. This probabilistic admission rule allows the server to maintain the customers' incentive compatibility in the least costly way. Since the difference in delay sensitivity is more significant than the difference in valuation, the server's main goal is to prevent customers with lower delay sensitivity from mimicking types with higher delay sensitivity. If the server solely admits type- $HH$  customers besides the  $HL$  and the  $LL$  types, the contract intended for type- $HH$  may look too appealing for the  $HL$  and  $LL$  types. In order to avoid this situation, the server may then be willing to allocate some capacity to serve the worst type (i.e., type- $LH$ ) customers.

If the server fully admits all customers, we find that only two priority classes are needed to differentiate between customers with high and low delay sensitivities. However, if at least one group of customers are admitted probabilistically, the server intends to offer higher priority to the  $HH$  type over the  $LH$  type, and he may randomize over the priority rule in order to achieve the best

incentive provision.

When the difference in valuation is relatively more significant (Cases 4b-1 to 4b-3), the server partially admits both the  $LL$  and  $LH$  type, to leverage information rent paid to customers with higher valuation. Similarly to Case 3b, the type- $HH$  customers may have the incentive to misreport as the  $LL$  and the  $HL$  type, and the server has to further differentiate between the  $LL$  and  $HL$  types with absolute or randomized priorities, in addition to the priority scheduling between the  $LH$  and  $HH$  types. Like the previously discussed cases, strategic idleness may also occur in an optimal solution, if the benefit outweighs the cost.

### 4.3 Novel Features of the Optimal Contracts

The aforementioned special cases reveal some interesting phenomena that arise from information asymmetry. For example, 1) the system may partially admit two customer groups, even though ex ante the server prefers one type to the other; 2) customer groups with the same delay sensitivities (but different valuations) may be awarded priority over one another; 3) instead of fully exhausting the system resource, it may be optimal to idle the server intentionally. Although we relegate the detailed derivations to the online appendix, here we provide intuitive explanations for these novel features.

#### 4.3.1 Mixed Admission Policy

We call an admission strategy a *mixed* policy, if it partially admits more than one customer types. Analogously, an *exclusive admission policy* is the one that partially admits at most one customer type. In Section 4.2, the mixed admission policy is applied in both Case 4a-1 and Case 4b-1.

Let us take Case 4a-1 and Case 3a as examples. In both cases, the type- $HL$  and type- $LL$  customers are fully admitted; however, the first case uses a mixed admission policy that partially admits both the type- $HH$  and type- $LH$  customers, while the admission policy for the second case is exclusive, with  $HH$  being the only partially admitted customer type. The main trade-off is as follows. In Case 4a-1,  $\Delta_v$ , the difference in valuation, is extremely small as compare to  $\Delta_c$ , the difference in delay sensitivity. The inclusion of type- $LH$  customers helps reduce the admission rate of the type- $HH$  customers while maintaining the same throughput level, which in turn helps to reduce the incentive of the type- $HL$  and type- $LL$  customers to misreport as the  $HH$  type. On

the other hand, this mixed admission strategy also results in a decrease in valuation, as well as an increase in the information rent to the type- $HH$  customer, both of which are determined by  $\Delta_v$ . Since the reduction in information rent of type- $HL$  and type- $LL$  customer is proportional to  $\Delta_c$  and is thus more significant than the decrease in valuation, the mixed admission policy can extract higher revenue for the server. Whereas in Case 3a,  $\Delta_v$  is not sufficiently small as compared to  $\Delta_c$ , and the cost of using the mixed admission policy outweighs the benefit. Therefore the exclusive admission policy is optimal in this case. The reasoning for the mixed admission policy in Case 4b-1 and the exclusive admission policy in Case 3b is similar.

To demonstrate the major trade-off, we illustrate the binding IC constraints of Case 4a-1 and Case 4b-1 in Figures 1 - 2. Here and in all the following diagrams, an edge from type- $ij$  to type- $i'j'$  represents the IC constraints that type- $i'j'$  does not be tempted to choose the contract intended for type- $ij$ . The value along an edge is the right-hand side of the corresponding IC constraint; in other words, it is a lower bound on the difference in the information rents received by the origin and destination customer types. The solid lines represent unique binding IC constraints, while the dotted lines imply multiple binding IC constraints.

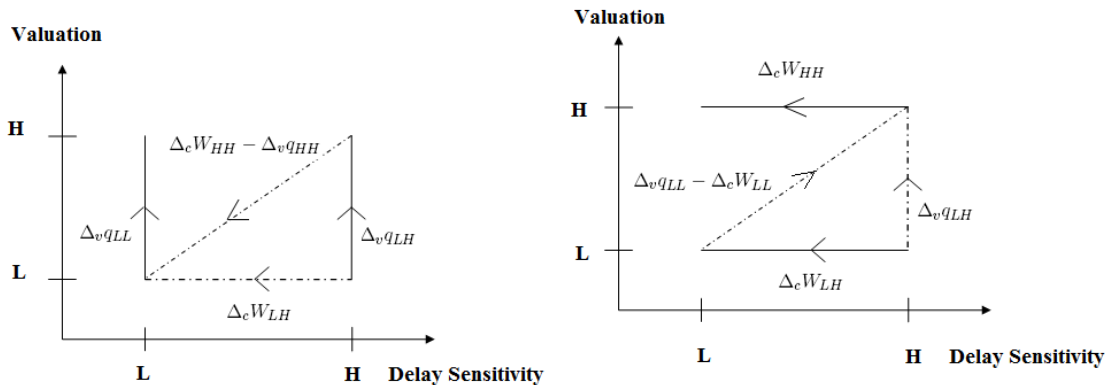


Figure 1: Binding IC constraints in Case 4a-1. Figure 2: Binding IC constraints in Case 4b-1.

### 4.3.2 Randomized Priority Scheduling

In a system under information symmetry, it is not necessary to differentiate customers with the same delay sensitivity but different valuations. Under information asymmetry, however, it may be optimal to give different priority rankings to these customers. For example, in Case 4a-2, the type-

$HH$  customers have higher service priority over the type- $LH$  customers. Moreover, the priority ranking between the  $HH$  and the  $LH$  is randomized to achieve the best information provision. Figure 3 shows the binding IC constraints in this case. In order to minimize the information rent of the type- $HL$  customers, the optimal menu of contracts needs to satisfy the following equation:

$$\Delta_c W_{LH} + \Delta_v q_{LL} = \Delta_v q_{LH} + \Delta_c W_{HH},$$

which can only be achieved through randomized priority ranking between the type- $LH$  and type- $HH$  customers.

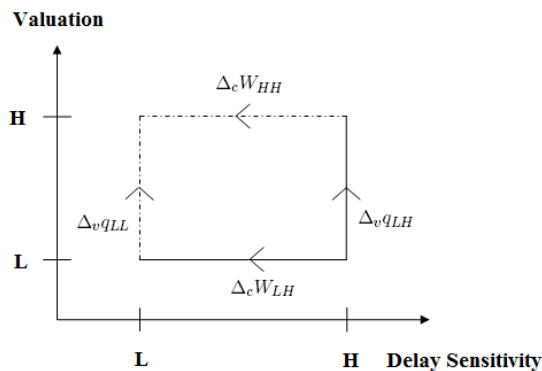


Figure 3: Binding IC constraints in Case 4a-2.

### 4.3.3 Strategic Idleness

If the server has complete knowledge of the customers' valuations and delay sensitivities, the optimal strategy is to exhaust the system resources, since any idleness will result in excessive costs. However, when the valuation and delay sensitivities are the customers' private information, it may be optimal to strategically delay the service for customers with lower delay sensitivities in order to provide appropriate incentives for customers with higher delay sensitivities. We call this non-exhaustion of system capacity the strategic idleness. This phenomena may arise in Case 3b, Case 4b-1 and Case 4b-2.

Let us take Case 3b as an example, in which the binding IC constraints are illustrated in Figure 4. If we increase  $W_{LL}$  by a small amount  $\epsilon > 0$ ,  $R_{HH}$ , the information rent received by type- $HH$

customers, will decrease by  $\Delta_c \epsilon$ , so will  $R_{HL}$  since it is equal to  $R_{HH} + \Delta_c W_{HH}$ . However, the delay cost will increase by  $c_L \epsilon$ . The benefit and cost trade-off depends on the relative significance of  $\Delta_c$  as compared to  $c_L$ , as well as the arrival rates  $\lambda_{LL}$ ,  $\lambda_{HH}$  and  $\lambda_{HL}$ . In situations where the benefit outweighs the cost, strategic idleness may emerge as an optimal solution.

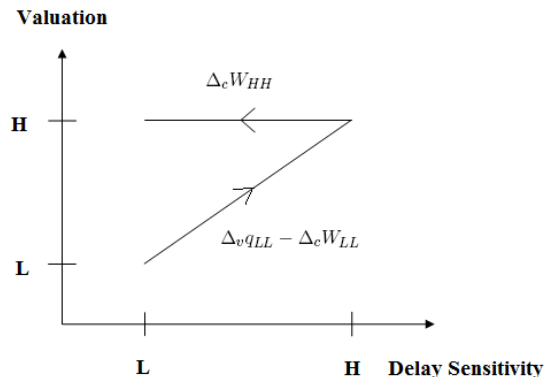


Figure 4: Binding IC constraints in Case 3b.

## 5 Revenue Gains from the Probabilistic Admission Policy

One of the key differences of our model and that of Afeche (2004) is the probabilistic admission control that allows the server to obtain more revenue from the customers. The revenue gains from a probabilistic admission policy follow from two sources. First, it helps the server to better utilize the limited resource; second, it provides the server the desired flexibility to leverage the information rents that must be paid to the customers. A natural question, that arises, is how much incremental revenue the probabilistic admission control can raise. To this end, we construct two examples to demonstrate the revenue gains from the probabilistic admission control policy, using the 0-1 (deterministic) admission policy (as in Afeche (2004)) as a benchmark. Example 5.1 demonstrates how the probabilistic admission policy gives the server the flexibility to accept the optimal level of work load, in order to balance the trade-off between the revenue gains from serving the customers and the delay costs due to congestion. Example 5.2 shows the revenue gains from leveraging the information rent by partially admitting two of the less favorable types ( $LH$  and  $LL$ ).

## 5.1 Example 1

In the first example, we use the following key parameter values:  $\mu = 100$ ,  $\lambda_{HL} = 30$ ,  $\lambda_{LL} = 30$ ,  $\lambda_{HH} = 10$ ,  $\lambda_{LH} = 10$ ,  $v_L = 5$ ,  $v_H = 6$ ,  $c_L = 100$ , and  $c_H = 300$ . Because  $v_H > c_L \frac{\mu}{(\mu - \lambda_{HH})^2}$ , the system should fully admit the type-*HL* (the most favorable) customers; i.e.,  $q_{HL} = 1$ . Also, because  $\frac{\Delta c}{\Delta v} > \mu(1 + \frac{\lambda_{HL}}{\lambda_{LL}})$ , the server always prefers type-*LL* customers to type-*HH* customers. Since  $v_L - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta v \frac{\lambda_{HL}}{\lambda_{LL}} < 0$ ,  $q_{HH} = q_{LH} = 0$ .

**Deterministic admission policy.** If the server uses the deterministic admission policy, then following the above argument we obtain that either  $q_{LL} = 0$  or  $q_{LL} = 1$ . If  $q_{LL} = 0$ , it follows that  $w_{HL} = \frac{1}{\mu - \lambda_{HL}} = \frac{1}{70}$ , and  $R_{HL} = 0$ . The server's revenue is:

$$\pi_s = v_H \lambda_{HL} - c_L \lambda_{HL} w_{HL} = 137.$$

On the other hand, if  $q_{LL} = 1$ ,  $w_{HL} = w_{LL} = \frac{1}{\mu - \lambda_{HL} - \lambda_{LL}} = \frac{1}{40}$ ,  $R_{LL} = 0$ , and  $R_{HL} = \Delta v q_{LL} = 1$ . The server's corresponding revenue is:

$$\pi_s = v_H \lambda_{HL} + v_L \lambda_{LL} - c_L (\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) - \lambda_{HL} R_{HL} = 150.$$

Thus, under the deterministic admission policy, the server's optimal revenue is 150.

**Probabilistic admission policy.** If instead the server adopts the probabilistic admission control policy, the optimal admission probability  $q_{LL}^*$  should be the solution to the following equation:

$$v_L - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} q_{LL}^*)^2} - \Delta v \frac{\lambda_{HL}}{\lambda_{LL}} = 0. \quad (9)$$

Solving (9) yields  $q_{LL}^* = \frac{2}{3}$ . It then follows that  $w_{HL}^* = w_{LL}^* = \frac{1}{\mu - \lambda_{HL} - \lambda_{LL} q_{LL}^*} = \frac{1}{50}$ ,  $R_{HL}^* = \Delta v q_{LL}^* = \frac{2}{3}$ , and the server's optimal revenue is:

$$\pi_s^* = v_H \lambda_{HL} + v_L \lambda_{LL} q_{LL}^* - c_L (\lambda_{HL} w_{HL}^* + \lambda_{LL} w_{LL}^*) - \lambda_{HL} R_{HL}^* = 160,$$

which is clearly higher than the highest revenue (150) achieved by the deterministic admission policy.

## 5.2 Example 2

In the second example, the following parameters are adopted:  $\mu = 100$ ,  $\lambda_{HL} = 30$ ,  $\lambda_{LL} = 20$ ,  $\lambda_{HH} = 10$ ,  $\lambda_{LH} = 20$ ;  $v_L = 20$ ,  $v_H = 21$ ,  $c_L = 100$ , and  $c_H = 400$ . In this case, since

$\frac{\Delta_c}{\Delta_v} > \mu(1 + \frac{\lambda_{HL}}{\lambda_{LL}})$ , the server always prefers type- $LL$  customers to type- $HH$  customers. Also, since  $v_L - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_v \frac{\lambda_{HL}}{\lambda_{LL}} > 0$ ,  $q_{HL} = q_{LL} = 1$ .

**Deterministic admission policy.** If the server uses the deterministic admission policy, she has three options: 1) admitting neither the type- $HH$  nor the type- $LH$  customers ( $q_{HH} = q_{LH} = 0$ ); 2) admitting only the type- $HH$  customers ( $q_{HH} = 1, q_{LH} = 0$ ); and 3) admitting both type- $HH$  and type- $LH$  customers ( $q_{HH} = q_{LH} = 1$ ). Her revenue associated with the three policies are calculated below.

When  $q_{HH} = q_{LH} = 0$ , it follows that  $w_{HL} = w_{LL} = \frac{1}{\mu - \lambda_{HL} - \lambda_{LL}} = \frac{1}{50}$ ,  $R_{LL} = 0$ , and  $R_{HL} = \Delta_v q_{LL} = 1$ . The server's corresponding revenue is:

$$\pi_s = v_H \lambda_{HL} + v_L \lambda_{LL} - c_L (\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) - \lambda_{HL} R_{HL} = 900.$$

In the second case where  $q_{HH} = 1, q_{LH} = 0$ , we obtain that  $w_{HH} = \frac{1}{\mu - \lambda_{HH}} = \frac{1}{90}$ , and  $w_{HL} = w_{LL} = \frac{1}{\lambda_{HL} + \lambda_{LL}} (\frac{\lambda_{HH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH}}{\mu - \lambda_{HH}}) = \frac{1}{36}$ . The information rents in this case are  $R_{HH} = 0$ ,  $R_{LL} = \Delta_c w_{HH} q_{HH} - \Delta_v q_{HH} = \frac{7}{3}$ , and  $R_{HL} = R_{LL} + \Delta_v q_{LL} = \frac{10}{3}$ . Accordingly, the server's revenue is:

$$\pi_s = v_H (\lambda_{HL} + \lambda_{HH}) + v_L \lambda_{LL} - c_L (\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) - c_H \lambda_{HH} w_{HH} - \lambda_{LL} R_{LL} - \lambda_{HL} R_{HL} = 910.$$

Finally, when  $q_{HH} = q_{LH} = 1$ ,  $w_{HH} = w_{LH} = \frac{1}{\mu - \lambda_{HH} - \lambda_{LH}} = \frac{1}{70}$ , and

$$w_{HL} = w_{LL} = \frac{1}{\lambda_{HL} + \lambda_{LL}} (\frac{\lambda_{HH} + \lambda_{LH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{LH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH} + \lambda_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}}) = \frac{1}{14}.$$

The information rents in this case are  $R_{LH} = 0$ ,  $R_{HH} = \Delta_v q_{LH} = 1$ ,  $R_{LL} = \Delta_c w_{LH} q_{LH} = \frac{30}{7}$ , and  $R_{HL} = R_{LL} + \Delta_v q_{LL} = \frac{37}{7}$ . The server's revenue is:

$$\begin{aligned} \pi_s &= v_H (\lambda_{HL} + \lambda_{HH}) + v_L (\lambda_{LL} + \lambda_{LH}) - c_L (\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) - c_H (\lambda_{HH} w_{HH} + \lambda_{LH} w_{LH}) \\ &\quad - \lambda_{HH} R_{HH} - \lambda_{LL} R_{LL} - \lambda_{HL} R_{HL} = 857. \end{aligned}$$

Collectively, the server's optimal revenue under the deterministic admission policy is 910.

**Probabilistic admission policy.** Now we investigate the case with the probabilistic admis-

sion policy. In such a scenario, the optimal contract should satisfy the following equations:

$$w_{HH} = \frac{1}{\mu - \lambda_{HH}q_{HH}}, \quad (10)$$

$$w_{LH} = \frac{1}{\lambda_{LH}q_{LH}} \left( \frac{\lambda_{HH}q_{HH} + \lambda_{LH}q_{LH}}{\mu - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH}} - \frac{\lambda_{HH}q_{HH}}{\mu - \lambda_{HH}q_{HH}} \right), \quad (11)$$

$$\Delta_c w_{LH}q_{LH} = \Delta_v q_{LH} + \Delta_c w_{HH}q_{HH} - \Delta_v q_{HH}, \quad (12)$$

$$v_L - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH})^2} - \Delta_c \frac{\mu}{(\mu - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH})^2} \left( 1 + \frac{\lambda_{HL} + \lambda_{LL}}{\lambda_{LH}} \right) - \Delta_v \frac{\lambda_{HH} + \lambda_{HL}}{\lambda_{LH}} = 0. \quad (13)$$

Solving (10)-(13) yields

$$q_{LH}^* = 0.3535, \quad q_{HH}^* = 0.4194, \quad w_{LH}^* = 0.0128, \quad w_{HH}^* = 0.0104.$$

It follows that

$$w_{HL}^* = w_{LL}^* = \frac{1}{\lambda_{HL} + \lambda_{LL}} \left( \frac{\lambda_{HH}q_{HH}^* + \lambda_{LH}q_{LH}^* + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH}q_{HH}^* - \lambda_{LH}q_{LH}^* - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH}q_{HH}^* + \lambda_{LH}q_{LH}^*}{\mu - \lambda_{HH}q_{HH}^* - \lambda_{LH}q_{LH}^*} \right) = 0.029,$$

and  $R_{LH}^* = 0$ ,  $R_{HH}^* = \Delta_v q_{LH}^* = 0.3535$ ,  $R_{LL}^* = \Delta_c w_{LH}^* q_{LH}^* = 1.3545$ , and  $R_{HL}^* = R_{LL}^* + \Delta_v q_{LL}^* = 2.3545$ . The server's optimal revenue is:

$$\begin{aligned} \pi_s^* &= v_H(\lambda_{HL} + \lambda_{HH}q_{HH}^*) + v_L(\lambda_{LL} + \lambda_{LH}q_{LH}^*) - c_L(\lambda_{HL}w_{HL}^* + \lambda_{LL}w_{LL}^*) - c_H(\lambda_{HH}w_{HH}^* + \lambda_{LH}w_{LH}^*) \\ &\quad - \lambda_{HH}R_{HH}^* - \lambda_{LL}R_{LL}^* - \lambda_{HL}R_{HL}^* = 962, \end{aligned}$$

which is clearly higher than that achieved by the best deterministic admission policy (910).

## 6 Conclusions

In this paper, we characterize the optimal joint pricing, scheduling, and admission control policy when the server faces customers with heterogeneous valuations and delay sensitivities. We show that the server always exhausts the most favorable type of customers (that have the highest valuation and are least sensitive to the delay) before admitting any other type of customers. Moreover, we find that even if the customers have identical valuations, in determining the admission policy, the server may still prefer one to another based on their delay sensitivity. Except the most favorable type of customers, the server's preference over the customers is endogenous. In particular, we find that the server may probabilistically admit more than one type.

We also characterize the optimal mechanisms in a number of special cases to gain further insights. Specifically, we find that the server may intend to offer different admission probabilities for the customers with common valuations, and may pool some groups of customers into one priority queue. Finally, occasionally a randomized priority may be adopted to prevent different types of customers to misrepresent themselves. Regarding the priority rules, we find that the server always optimally gives the customers with high delay sensitivity the absolute priority over those with low delay sensitivity. Moreover, to distinguish among different groups of customers, it may be the server's best interest to insert some strategic idleness and use randomized priority rules.

As we intend to provide a simple framework to illustrate the above managerial implications, our stylized model certainly has its own limitations and may be extended in various dimensions. In this paper, we focus exclusively on the case when the server makes a one-time decision on managing his business. In reality, there might be situations in which this can be done dynamically. For example, if the server is able to adjust dynamically the admission control based on the current queueing status, he may be able to strategically select the appropriate customers to serve. Also, if the scheduling policy can be adjusted over time, if the server attempts to serve a specific type of customers, he may be able to (temporarily) give the highest priority to those customers, thereby reducing their disutility that arises from the congestion. Extending our framework to a dynamic setting is a crucial step as well as a challenging task.

Another possible direction is to consider nonlinear delay cost. In contrast with our current setting in which only the expected delay is active, higher moments of the delay may also affect the customers' utility in such a nonlinear environment (Yahalom et al. (2005)). In this scenario, while designing the priority rule, the server may also intend to minimize, for example, the variance of the delay a customer may encounter. In particular, static priority rule may be suboptimal (Yahalom et al. (2005)) and one inevitably needs to search among those dynamic scheduling policy (e.g., the generalized  $c\mu$  rule). Including these effects may broaden the applicability of our framework, and it definitely deserves further investigation.

## References

Afeche, P. 2004. Incentive compatible revenue management in queueing systems: Optimal strategic idleness and other delay tactics. Working paper, Northwestern University.

- Armstrong, M. 1996. Multiproduct nonlinear pricing. *Econometrica* **64**(1) 51–75.
- Armstrong, M. 2000. Optimal multi-object auctions. *Review of Economic Studies* **67** 455–481.
- Armstrong, M., J. Rochet. 1999. Multi-dimensional screening: A user’s guide. *European Economic Review* **43** 959–979.
- Asker, J., E. Cantillon. 2009. Procurement when price and quality matter. Forthcoming in RAND Journal of Economics.
- Cachon, G., M. Lariviere. 1999. Capacity choice and allocation: strategic behavior and supply chain performance. *Management Science* **45** 1091–1108.
- Chang, X., D. Petr. 2001. A survey of pricing for integrated service networks. *Computer Communications* **24** 1808–1818.
- Coffman, E., I. Mitrani. 1980. A characterization of waiting time performance realizable by single-server queues. *Operations Research* **28** 810–821.
- Corbett, C., X. de Groot. 2000. A supplier’s optimal quantity discount policy under asymmetric information. *Management Science* **46** 444–450.
- Gibbens, R., F. Kelly, P. Key. 1995. A decision-theoretic approach to call admission control in ATM networks. *IEEE Journal of Selected Areas in Communications* **13** 1101–1114.
- Ha, A. 2001. Supplier-buyer contracting: Asymmetric cost information and cutoff level policy for buyer participation. *Naval Research Logistics* **48** 41–64.
- Hassin, R., M. Haviv. 2002. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.
- Iyer, A., L. Schwarz, S. Zenios. 2005. A principal-agent model for product specification and production. *Management Science* **51** 106–119.
- Katta, A.-K., J. Sethuraman. 2005. Pricing strategies and service differentiation in queues - a profit maximization perspective. Working paper, Columbia University.
- Laffont, J., D. Martimort. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, USA.

- Lewis, J., R. Russell, F. Toomey, B. Mcgurk, S. Crosby, I. Leslie. 1998. Practical connection admission control for atm networks based on on-line measurements. *Computer Communications* **21** 1585–1596.
- Lutze, H., O. Ozer. 2008. Promised lead time contracts under asymmetric information. Forthcoming in *Operations Research*.
- Mendelson, H. 1985. Pricing computer services: Queueing effects. *Communications of ACM* **28** 312–321.
- Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research* **38** 870–883.
- Naor, P. 1969. On the regulation of queue size by levying tolls. *Econometrica* **37** 15–24.
- Rochet, J., L. Stole. 2005. The economics of multidimensional screening. M. Dewatripont, L. Hansen, S. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications - Eight World Congress*. Econometric Society Monographs, Cambridge University Press, New York.
- Shanthikumar, J., D. Yao. 1992. Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research* **40** 293–299.
- Stidham, S. 1985. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* **30** 705–713.
- Stidham, S. 2002. Analysis, design and control of queueing systems. *Operations Research* **50** 197–216.
- Wilson, R. 1993. *Nonlinear Pricing*. Oxford University Press, New York.
- Wu, S.-Y., P.-Y. Chen. 2008. Versioning and piracy control for digital information goods. Forthcoming in *Operations Research*.
- Yahalom, T., J. M. Harrison, S. Kumar. 2005. Designing and pricing incentive compatible grades of service in queueing systems. Working paper 1032, Stanford University.
- Yang, Z., G. Aydin, V. Babich, D.R. Beil. 2009. Supply disruptions, asymmetric information and a backup production option .

Zhang, H., S. Zenios. 2008. A dynamic principal-agent model with hidden information: Sequential optimality through truthful state revelation. Forthcoming in *Operations Research*.

Zhang, L., S. Deering, D. Estrin, S. Shenker. 1993. RSVP: A new resource reservation protocol. *IEEE Network* **7** 8–18.

## Appendix. Proof of Lemmas and Propositions

Before we prove the propositions, we first introduce the following technical lemmas that are necessary for establishing our results.

**Lemma 1.** *Let  $(q^1, W^1, R)$  and  $(q^2, W^2, R)$  be two solutions such that for some  $ij \in T$ ,*

$$\begin{aligned} q_{ij}^2 &\geq q_{ij}^1, \text{ if } ij \in \{HL, HH\}, \\ q_{ij}^2 &\leq q_{ij}^1, \text{ if } ij \in \{LL, LH\}, \\ W_{ij}^2 &\leq W_{ij}^1, \text{ if } ij \in \{LH, HH\}, \\ W_{ij}^2 &\geq W_{ij}^1, \text{ if } ij \in \{LL, HL\}. \end{aligned}$$

*If  $(q^1, W^1, R)$  satisfies constraint  $IC(i'j' - ij)$ , then  $(q^2, W^2, R)$  also satisfies it.*

*Proof.* Consider the following constraint  $IC(i'j' - ij)$ :

$$R_{i'j'} - R_{ij} \geq (v_{i'} - v_i)q_{ij} - (c_{j'} - c_j)W_{ij}.$$

Since the left hand side of the constraint only depends on  $R$ , if we could show that

$$(v_{i'} - v_i)q_{ij}^2 - (c_{j'} - c_j)W_{ij}^2 \leq (v_{i'} - v_i)q_{ij}^1 - (c_{j'} - c_j)W_{ij}^1, \quad \forall ij, i'j' \in T, \quad (14)$$

then  $(q^2, W^2, R)$  must satisfy this constraint if  $(q^1, W^1, R)$  does. To see this is the case, we first consider  $ij = LH$ . In this case,  $v_{i'} - v_i \geq 0$  and  $c_{j'} - c_j \leq 0$ , since  $v_i = v_L$  and  $c_j = c_H$ . Because  $q_{ij}^2 \leq q_{ij}^1$  and  $W_{ij}^2 \leq W_{ij}^1$ , (14) must hold. Following similar procedures, one can show that (14) holds if  $ij$  is  $LL$ ,  $HH$  or  $HL$ .  $\square$

**Lemma 2.**  $W_{LH} \leq W_{LL}$ , and  $W_{HH} \leq W_{HL}$ .

*Proof.* From the IC constraints  $IC(LL - LH)$  and  $IC(LH - LL)$ , we have  $R_{LL} - R_{LH} \geq \Delta_c W_{LH}$ , and  $R_{LH} - R_{LL} \geq -\Delta_c W_{LL}$ . Adding the above two inequalities together, it follows that  $0 \geq \Delta_c (W_{LH} - W_{LL})$ , which implies that  $W_{LH} \leq W_{LL}$ . Similarly, one can verify that  $IC(HL - HH)$  and  $IC(HH - HL)$  implies  $W_{HH} \leq W_{HL}$ .  $\square$

**Proof of Proposition 4.** We divided the proof in three parts, each dedicated to show that if the optimal solution does not fully admit the type- $HL$  customers, then it is optimal not to admit any the

type- $LH$ ,  $LL$  or  $HH$  customers respectfully. The proof is by contradiction. Thus, in the following three parts we posit hypotheses by negation and then show that they result in contradictions.

*Part 1:  $q_{LH}^* > 0$  and  $q_{HL}^* < 1$ .*

In this case, for some positive number  $\epsilon$ , we construct a new solution  $(q', W', R')$  as follows:

$$\begin{aligned} q'_{ij} &= \begin{cases} q_{ij}^* + \frac{\epsilon}{\lambda_{ij}} & ij = HL \\ q_{ij}^* - \frac{\epsilon}{\lambda_{ij}} & ij = LH \\ q_{ij}^* & ij \in \{HH, LL\} \end{cases}, \\ W' &= W^*, \\ R' &= R^*. \end{aligned}$$

By Lemma 1, the new solution satisfies all IC constraints. We only need to check the resource constraint  $RE(S)$  for all  $S \subseteq T$ . If  $S = \{HL\}$ , then  $RE(S)$  is not binding at  $(q^*, W^*, R^*)$ , because from Proposition 6  $RE(\{LH, HH\})$  is binding, leading to a conflict against the properties of a polymatroid. In this case, we can always find an  $\epsilon$  small enough so that  $RE(\{HL\})$  is satisfied at  $(q', W', R')$ . On the other hand, if  $S \neq \{HL\}$ , the right-hand side of  $RE(S)$  never increases in the new solution, while the left-hand side remains unchanged. Therefore, the new solution satisfies all resource constraints.

In summary, the new solution is feasible, and the objective increases by  $\Delta_v \epsilon$  as compared to  $(q^*, W^*, R^*)$ . This leads to a contradiction to the optimality of  $(q^*, W^*, R^*)$ .

*Part 2:  $q_{LL}^* > 0$  and  $q_{HL}^* < 1$ .*

In this case, for some  $\epsilon > 0$ , we define a new mechanism  $(q', W', R')$  as follows:

$$\begin{aligned} q'_{ij} &= \begin{cases} q_{ij}^* + \frac{\epsilon}{\lambda_{ij}} & ij = HL \\ q_{ij}^* - \frac{\epsilon}{\lambda_{ij}} & ij = LL \\ q_{ij}^* & ij \in \{HH, LL\} \end{cases}, \\ W' &= W^*, \\ R' &= R^*. \end{aligned}$$

Following similar procedures as in Case 1, one can verify that  $(q', W', R')$  satisfies all IC and resource constraints for sufficiently small  $\epsilon$ . Furthermore, the objective value under this new solution increases by  $\Delta_v \epsilon$ , thus leading to a contradiction to the fact that  $(q^*, W^*, R^*)$  is optimal.

Part 3:  $q_{HH}^* > 0$  and  $q_{HL}^* < 1$ .

Suppose that  $q_{HH}^* > 0$  and  $q_{HL}^* < 1$ . From parts 1 and 2, we know that  $q_{LH}^* = q_{LL}^* = 0$ . In this case,  $RE(\{HH\})$  and  $RE(\{HH, HL\})$  must be binding, otherwise the solution would be suboptimal. This implies  $W_{HH}^* = \frac{q_{HH}^*}{\mu - \lambda_{HH}q_{HH}^*}$ , and

$$W_{HL}^* = \frac{q_{HL}^* + \frac{\lambda_{HH}}{\lambda_{HL}}q_{HH}^*}{\mu - \lambda_{HL}q_{HL}^* - \lambda_{HH}q_{HH}^*} - \frac{q_{HH}^*}{\mu - q_{HH}^*\lambda_{HH}}.$$

To disprove the optimality of  $(q^*, W^*, R^*)$ , we define a new solution  $(q', W', R')$  as below:

$$\begin{aligned} q'_{ij} &= \begin{cases} q_{ij}^* + \frac{\epsilon}{\lambda_{ij}} & ij = HL \\ q_{ij}^* - \frac{\epsilon}{\lambda_{ij}} & ij = HH \\ 0 & ij \in \{LH, LL\} \end{cases}, \\ W'_{ij} &= \begin{cases} \frac{q'_{HL} + \frac{\lambda_{HH}}{\lambda_{HL}}q'_{HH}}{\mu - \lambda_{HL}q'_{HL} - \lambda_{HH}q'_{HH}} - \frac{q'_{HH}}{\mu - \lambda_{HH}q'_{HH}} & ij = HL \\ \frac{q'_{HH}}{\mu - \lambda_{HH}q'_{HH}} & ij = HH \\ 0 & ij \in \{LH, LL\}. \end{cases}, \\ R' &= R^*. \end{aligned}$$

By construction,  $\lambda_{HH}W'_{HH} + \lambda_{HL}W'_{HL} = \lambda_{HH}W_{HH}^* + \lambda_{HL}W_{HL}^*$ , and it can be verified that the objective function,  $\sum_{ij \in T} \lambda_{ij}(v_i q_{ij} - c_j W_{ij} - R_{ij})$ , increases by a positive amount of  $\Delta_c(\lambda_{HH}W_{HH}^* - \lambda_{HH}W'_{HH})$  in the new solution. As all the resource constraints are clearly satisfied by the new solution, we just need to check the IC constraints. Following Lemma 1,  $IC(ij - HL)$  is satisfied, since  $q'_{HL} > q_{HL}^*$  and  $W'_{HL} > W_{HL}^*$ . Furthermore,  $IC(LH - HH)$  and  $IC(HL - HH)$  can not be binding at  $(q^*, W^*, R^*)$ , since  $R_{HH}^* = 0$ , which means the left-hand sides of these constraints are nonnegative, while the right-hand sides are negative. Therefore the small changes in  $q_{HH}$  and  $W_{HH}$  will not affect these constraints.

The only constraint that might be violated by the new solution is  $IC(LL - HH)$ , and this only happens when  $IC(LL - HH)$  is binding at  $(q^*, W^*, R^*)$ . Suppose this is the case, define  $f(q_{HH})$  to be the right-hand side of  $IC(LL - HH)$ :

$$f(q_{HH}) \equiv \Delta_c \frac{q_{HH}}{\mu - \lambda_{HH}q_{HH}} - \Delta_v q_{HH}.$$

The derivative of  $f$  is  $f'(q_{HH}) = \frac{\mu}{(\mu - \lambda_{HH}q_{HH})^2}$ , which is increasing in  $q_{HH}$ . This suggests that  $f$  is modular. Because  $f(0) = 0$ ,  $f(q_{HH}^*) = R_{LL}^* - R_{HH}^* \geq 0$ , and  $0 \leq q'_{HH} < q_{HH}^*$ , it must be true that

$f(q'_{HH}) \leq f(q^*_{HH})$ , which implies  $IC(LL - HH)$  is not violated by the new solution.  $\square$

**Proof of Proposition 5.** First, recall the incentive compatibility constraints  $IC(HH - LH)$  and  $IC(LH - HH)$ ,  $R^*_{HH} - R^*_{LH} \geq \Delta_v q^*_{LH}$  and  $R^*_{LH} - R^*_{HH} \geq -\Delta_v q^*_{HH}$ . Adding the above two inequalities together, it follows that  $0 \geq \Delta_v (q^*_{LH} - q^*_{HH})$ , which implies that  $q^*_{LH} \leq q^*_{HH}$ . Similarly, one can verify that  $IC(HL - LL)$  and  $IC(LL - HL)$  implies  $q^*_{LL} \leq q^*_{HL}$ .

Second, the claim  $q^*_{LL} \leq q^*_{HL}$  follows straightly from Proposition 4, since either  $q^*_{HH} = 0$  or  $q^*_{HL} = 1$ .

Finally, we show  $q^*_{LH} \leq q^*_{LL}$  by contradiction. Suppose that  $q^*_{LH} > q^*_{LL}$ , we construct a new solution  $(q', W', R')$  as below

$$\begin{aligned} q'_{ij} &= \begin{cases} q^*_{LH} - \frac{\epsilon}{\lambda_{LH}} & ij = LH \\ q^*_{LL} + \frac{\epsilon}{\lambda_{LL}} & ij = LL \\ q^*_{ij} & ij \in \{HH, HL\} \end{cases}, \\ W'_{ij} &= \begin{cases} W^*_{LH} - \frac{\epsilon}{\lambda_{LH}} \frac{1}{\mu - \lambda_{LH} q^*_{LH}} & ij = LH \\ W^*_{LL} + \frac{\epsilon}{\lambda_{LL}} \frac{1}{\mu - \lambda_{LH} q^*_{LH}} & ij = LL \\ W^*_{ij} & ij \in \{HH, HL\} \end{cases}, \\ R' &= R^*. \end{aligned}$$

The objective value increased by  $\Delta_v \frac{\epsilon}{\mu - \lambda_{LH} q^*_{LH}}$  under the new solution. Next we show that it satisfy all IC and resource constraints. Among all resource constraints, we limit our attention to  $RE(\{LH\})$  and  $RE(\{LH, HH\})$ , All resource constraints are either non-binding at  $(q^*, W^*, R^*)$ , or not affected by the change. To see  $RE(\{LH\})$  still holds under the new solution:

$$\begin{aligned} \lambda_{LH} W'_{LH} &= \lambda_{LH} W^*_{LH} - \frac{\epsilon}{\mu - \lambda_{LH} q^*_{LH}} \\ &\geq \frac{\lambda_{LH} q^*_{LH}}{\mu - \lambda_{LH} q^*_{LH}} - \frac{\epsilon}{\mu - \lambda_{LH} q^*_{LH}} \\ &= \frac{\lambda_{LH} q'_{LH}}{\mu - \lambda_{LH} q^*_{LH}} \\ &\geq \frac{\lambda_{LH} q'_{LH}}{\mu - \lambda_{LH} q'_{LH}}. \end{aligned}$$

In addition, it follows that

$$\begin{aligned}
& \lambda_{LH}W'_{LH} + \lambda_{HH}W'_{HH} = \lambda_{LH}W^*_{LH} + \lambda_{HH}W^*_{HH} - \frac{\epsilon}{\mu - \lambda_{LH}q^*_{LH}} \\
& \geq \frac{\lambda_{LH}q^*_{LH} + \lambda_{HH}q^*_{HH}}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} - \frac{\epsilon}{\mu - \lambda_{LH}q^*_{LH}} \\
& \geq \frac{\lambda_{LH}q^*_{LH} + \lambda_{HH}q^*_{HH}}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} - \frac{\epsilon}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} \\
& = \frac{\lambda_{LH}q'_{LH} + \lambda_{HH}q'_{HH}}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} \\
& \geq \frac{\lambda_{LH}q'_{LH} + \lambda_{HH}q'_{HH}}{\mu - \lambda_{LH}q'_{LH} - \lambda_{HH}q'_{HH}},
\end{aligned}$$

which implies that  $RE(\{LH, HH\})$  is satisfied by the new solution.

Following from Lemma 1, all IC constraints remain valid under the new solution except for  $IC(HH - LL)$  and  $IC(HL - LL)$ . Our next step is to show that both of them are non-binding at  $(q^*, W^*, R^*)$ .

By contradiction, if  $IC(HH - LL)$  is binding at  $(q^*, W^*, R^*)$ , it follows that

$$R^*_{HH} = R^*_{LL} + \Delta_v q^*_{LL} - \Delta_c W^*_{LL} < R^*_{LL} - \Delta_c W^*_{LL} + \Delta_v q^*_{LH} \leq R^*_{LH} + \Delta_v q^*_{LH},$$

where the first inequality follows from the assumption that  $q^*_{LL} < q^*_{LH}$ , and the second one follows from  $IC(LH - LL)$ . Clearly, this leads to a contradiction to  $IC(HH - LH)$ .

Additionally, if  $IC(HL - LL)$  is binding, we claim that either  $IC(LL - LH)$  or  $IC(LL - HH)$  must also binding. First of all, at least one of  $IC(LL - LH)$ ,  $IC(LL - HH)$  and  $IC(LL - HL)$  has to be binding, otherwise  $(q^*, W^*, R^*)$  is suboptimal. Secondly,  $IC(LL - HL)$  cannot be binding when  $IC(HL - LL)$  is binding, due to the fact that  $q^*_{LL} < q^*_{LH} \leq q^*_{HL}$ .

If both  $IC(HL - LL)$  and  $IC(LL - LH)$  are binding, it follows that

$$R^*_{HL} = R^*_{LL} + \Delta_v q^*_{LL} = R^*_{LH} + \Delta_c W^*_{LH} + \Delta_v q^*_{LL} < R^*_{LH} + \Delta_c W^*_{LH} + \Delta_v q^*_{LH},$$

which leads to a contradiction to  $IC(HL - LH)$ .

On the other hand, if both  $IC(HL - LL)$  and  $IC(LL - HH)$  are binding, it follows that

$$R^*_{HL} = R^*_{LL} + \Delta_v q^*_{LL} = R^*_{HH} + \Delta_c W^*_{HH} - \Delta_v q^*_{HH} + \Delta_v q^*_{LL} < R^*_{HH} + \Delta_c W^*_{HH},$$

where the inequality follows from the fact that  $q^*_{LL} < q^*_{LH} \leq q^*_{HH}$ . Clearly, this leads to a contradiction to  $IC(HL - HH)$ .  $\square$

**Proof of Proposition 6.** Suppose that  $RE(\{LH, HH\})$  is not binding at  $(q^*, W^*, R^*)$ , then at most one of  $RE(\{LH\})$  and  $RE(\{HH\})$  can be binding. Because from the properties of a polymatroid, if both  $RE(\{LH\})$  and  $RE(\{HH\})$  are binding, it must be that  $q_{LH}^* = q_{HH}^* = 0$ , which implies that  $RE(\{LH, HH\})$  is binding. Let  $i_1j_1 \in \{LH, HH\}$  be a type such that  $RE(\{ij\})$  is not binding. We consider the following two cases, depending on whether there exists a set  $S \subseteq T$  such that  $RE(S)$  is binding at  $(q^*, W^*, R^*)$  and  $LH \in S$  (Case 1) or not (Case 2).

*Case 1:*

Suppose we can find a set  $S \subseteq T$  such that  $RE(S)$  is binding at  $(q^*, W^*, R^*)$  and  $LH \in S$ . We let  $S_0$  to be the minimal among such sets. It must be true that  $S_0 \cap \{LL, HL\} \neq \Phi$ , since otherwise  $S_0 = \{LH, HH\}$ , which means  $RE(\{LH, HH\})$  is binding. Let  $i_2j_2 \in S_0 \cap \{LL, HL\}$ . We construct a new solution  $(q', W', R')$  as follows:

$$\begin{aligned} q' &= q^*, \\ W'_{ij} &= \begin{cases} W_{ij}^* - \frac{\epsilon}{\lambda_{ij}} & ij = i_1j_1 \\ W_{ij}^* + \frac{\epsilon}{\lambda_{ij}} & ij = i_2j_2 \\ W_{ij}^* & ij \in T \setminus \{i_1j_1, i_2j_2\}, \end{cases} \\ R' &= R^*, \end{aligned}$$

where  $\epsilon > 0$  is a sufficiently small number.

From Lemma 1, the new solution satisfies all IC constraints, we only need to it also satisfies the resource constraint  $RE(S)$  for all  $S \subseteq T$ . As a reminder, the resource constraint is formulated as  $\sum_{ij \in S} \lambda_{ij} W_{ij} \geq \frac{\sum_{ij \in S} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in S} \lambda_{ij} q_{ij}}$ .

Clearly,  $RE(S)$  is satisfied if  $S_0 \subseteq S$ , since both sides of the constraint remain unchanged under the new solution, as compare to  $(q^*, W^*, R^*)$ . If  $S \subset S_0$ , but  $i_1j_1 \notin S$ ,  $RE(S)$  is not violated by the new solution, since its left-hand side can only increase, while its right-hand side remains the same. Finally, if  $S \subset S_0$  and  $i_1j_1 \in S$ , it must be true that  $RE(S)$  is not binding at  $(q^*, W^*, R^*)$ . In this case, we could always find an  $\epsilon$  small enough so that the new solution does not violate  $RE(S)$ .

However, the server's revenue,  $\sum_{ij \in T} \lambda_{ij}(v_i q_{ij} - c_j W_{ij} - R_{ij})$ , increases by  $\Delta_c \epsilon$  in the new solution, which leads to a contradiction to the optimality of  $(q^*, W^*, R^*)$ .

*Case 2:*

If we cannot find a set  $S \subseteq T$  that contains the type- $ij$  such that  $RE(S)$  is binding at

$(q^*, W^*, R^*)$ , for some  $\epsilon > 0$ , we construct a new solution  $(q', W', R')$  as below:

$$\begin{aligned} q' &= q^*, \\ W'_{ij} &= \begin{cases} W_{ij}^* - \epsilon & ij = i_1j_1 \\ W_{ij}^* & ij \in T \setminus \{i_1j_1\} \end{cases}, \\ R' &= R^*. \end{aligned}$$

This new solution clearly yields a higher revenue for the service provider. It also satisfies all the IC constraints, following Lemma 1. If we choose a small enough  $\epsilon$ , the new solution also satisfy the resource constraint  $RE(S)$  for all  $S \subseteq T$ , because either  $RE(S)$  is not binding at  $(q^*, W^*, R^*)$ , or  $S$  does not contain type- $i_1j_1$ . This again leads to a contradiction to the fact that  $(q^*, W^*, R^*)$  is optimal.  $\square$

**Proof of Proposition 7.** Without loss of generality, we assume that  $q_{LH}^* > 0$ , because if  $q_{LH}^* = 0$ ,  $RE(\{HH\})$  is equivalent to  $RE(\{LH, HH\})$ , which is binding by Proposition 6. Suppose  $q_{HH}^* < 1$ , but  $RE(\{HH\})$  is non-binding. For some  $\epsilon > 0$ , we construct a new solution  $(q', W', R')$  as follows:

$$\begin{aligned} q'_{ij} &= \begin{cases} W_{LH}^* - \frac{\epsilon}{\lambda_{LH}} & ij = LH \\ W_{HH}^* + \frac{\epsilon}{\lambda_{LH}} & ij = HH \\ W_{ij}^* & ij \in \{LL, HL\} \end{cases}, \\ W' &= W^*, \\ R' &= R^*. \end{aligned}$$

By Lemma 1, all IC constraints are satisfied by the new solution. All resource constraints must also be satisfied, except for  $RE(\{HH\})$ . Since  $RE(\{HH\})$  is non-binding at  $(q^*, W^*, R^*)$ , we can always find an  $\epsilon$  small enough so that this resource constraint is not violated by the new solution. However, the object value increase by  $\Delta_v \epsilon$  under the new solution, leading to a contradiction to the optimality of  $(q^*, W^*, R^*)$ .  $\square$

## Online Appendix for

# “Pricing, scheduling, and admission control in queueing systems: A mechanism design approach Detailed Solutions to the Special Cases

In this online appendix, we construct detailed solutions to some special cases. All of proofs follow three identical steps. In *Step 1*, we construct an upper bound function  $g(q)$  or  $g(q, W)$  of the server’s objective  $Z(q, W, R)$  by replacing all the information rent  $R$  and the expected delay  $W$  of some customer types with their valid lower bounds. In *Step 2*, we then search for  $\tilde{q}$  or  $(\tilde{q}, \tilde{W})$ , the optimal solution to  $\max_q \{g(q) : 0 \leq q_{ij} \leq 1 \forall ij \in T\}$  or  $\max_{q, W} \{g(q, W) : 0 \leq q_{ij} \leq 1 \forall ij \in T, (RE)\}$ . Based on the concavity of  $g(q)$  or  $g(q, W)$ , it suffice to look for solutions that satisfy the first order conditions. In *Step 3*, we construct a feasible solution  $(q^*, W^*, R^*)$  where  $q^*$  equals  $\tilde{q}$ , and  $W^*$  and  $R^*$  are set to their lower bounds used in *Step 1*. Then we verify that  $Z(q^*, W^*, R^*)$  is equal to  $g(\tilde{q})$  or  $g(q, \tilde{W})$ , an upper bound on the server’s maximum revenue, validating  $(q^*, W^*, R^*)$  as an optimal solution to the server’s problem. In the sequel we present our results as propositions and prove them accordingly following the aforementioned procedures.

### One Group of Customers

**Proposition 8.** *If  $v_H - \frac{c_L}{(\mu - \lambda_{HL})^2} \leq 0$ , the optimal mechanism is to admit only the type- $HL$  customers.*

*Proof. Step 1:* Constructing an upper bound function.

The server’s objective function satisfies

$$\begin{aligned}
 Z(q, W, R) &\equiv \sum_{ij \in T} \lambda_{ij} (v_i q_{ij} - c_j W_{ij} - R_{ij}) \\
 &= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
 &\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} \\
 &\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} \\
 &\equiv g(q),
 \end{aligned}$$

where the first inequality is due to the nonnegativity of  $q$ ,  $W$  and  $R$ , and the second follows from  $RE(T)$ . Note that in this case the upper bound function only depends on  $q$ .

*Step 2:* Optimizing the upper bound function.

Define  $\tilde{q}_{LH} = \tilde{q}_{LL} = \tilde{q}_{HH} = 0$ , and  $\tilde{q}_{HL}$  to be the solution of  $v_H - \frac{c_L}{(\mu - \lambda_{HL}q_{HL})^2} = 0$ . Because  $v_H - \frac{c_L}{(\mu - \lambda_{HL})^2} \leq 0$ ,  $\tilde{q}_{HL} \leq 1$ . In addition,

$$\frac{\partial g}{\partial q_{ij}}|_{q=\tilde{q}} = \lambda_{ij}(v_H - \frac{c_L}{(\mu - \lambda_{HL}\tilde{q}_{HL})^2}) = 0, \quad \forall ij \in T.$$

Since  $g$  is concave in  $q$ ,  $\tilde{q}$  is optimal to the optimization problem  $\max_q \{g(q) : 0 \leq q_{ij} \leq 1 \forall ij \in T\}$ .

*Step 3:* Constructing a feasible solution.

Given  $\tilde{q}$ , we construct a solution  $(q^*, W^*, R^*)$  as below

$$\begin{aligned} q_{ij}^* &= \begin{cases} \frac{y^*}{\lambda_{HL}} & ij = HL \\ 0 & ij \in \{LH, LL, HH\} \end{cases}, \\ W_{ij}^* &= \begin{cases} \frac{q_{HL}^*}{\mu - \lambda_{HL}q_{HL}^*} & ij = HL \\ 0 & ij \in \{LH, LL, HH\} \end{cases}, \\ R_{ij}^* &= 0, \quad \forall ij \in T. \end{aligned}$$

Since  $Z(q^*, W^*, R^*) = g(\tilde{q})$ , it is clear an upper bound on any feasible solution. We only need to verify that  $(q^*, W^*, R^*)$  is feasible, which is obvious since all IC and resource constraints are satisfied.  $\square$

## Two Groups of Customers

**Proposition 9.** *The optimal mechanism is to fully admit the HL and partially admit the type-LL customers with equal priority, if the following conditions hold:*

- $\Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \leq \Delta_c \frac{1}{\mu}$ ,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) > 0$ ,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \leq 0$ .

*Proof. Step 1:* Constructing an upper bound function.

$$\begin{aligned}
& Z(q, W, R) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v \lambda_{LL} q_{LL} - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \lambda_{HL} \Delta_v q_{LL} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \Delta_v \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) \lambda_{LL} q_{LL} - \Delta_c \frac{\lambda_{LH} q_{LH} + \lambda_{HH} q_{HH}}{\mu - \lambda_{LH} q_{LH} - \lambda_{HH} q_{HH}} \\
&\equiv g(q).
\end{aligned}$$

Here the first inequality is due to  $IC(HL - LL)$  and the nonnegativity of  $q$  and  $R$ ; the second follows from  $RE(T)$  and  $RE(\{LH, HH\})$ .

*Step 2:* Optimizing the upper bound function.

Define  $\tilde{q}_{HL} = 1$ ,  $\tilde{q}_{LH} = \tilde{q}_{HH} = 0$ , and  $\tilde{q}_{LL}$  to be the solution to

$$h(q_{LL}) \equiv v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} q_{LL})^2} - \Delta_v \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) = 0.$$

We claim that  $\tilde{q}$  is the optimal solution to  $\max_q \{g(q) : 0 \leq q_{ij} \leq 1, \forall ij \in T\}$ .

By assumption, we have

$$\begin{aligned}
h(0) &= v_H - \frac{c_L}{(\mu - \lambda_{HL})^2} - \Delta_v \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) > 0, \\
h(1) &= v_H - \frac{c_L}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_v \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) \leq 0.
\end{aligned}$$

Because  $h(q_{LL})$  is decreasing in  $q_{LL}$ , it must be true that  $0 < \tilde{q}_{LL} \leq 1$ .

We now claim that  $\tilde{q}$  satisfies the KKT (Karush-Kuhn-Tucker) conditions of the optimization problem  $\max_q \{g(q) : 0 \leq q_{ij} \leq 1, \forall ij \in T\}$ . Since  $g(q)$  is a concave function,  $\tilde{q}$  must be an optimal solution to this problem and  $g(\tilde{q})$  is an upper bound on  $Z(q, W, R)$ . To see that the KKT

conditions are satisfied, we observe that

$$\begin{aligned}
& \frac{\partial g}{\partial q_{LL}}|_{q=\tilde{q}} = \lambda_{LL}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
& = \lambda_{LL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
& = 0, \\
& \frac{\partial g}{\partial q_{HL}}|_{q=\tilde{q}} = \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2}] \\
& = \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2}] \\
& > \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
& = 0, \\
& \frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} = \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_c \frac{\mu}{(\mu - \lambda_{LH} \tilde{q}_{LH} - \lambda_{HH} \tilde{q}_{HH})^2}] \\
& = \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_c \frac{1}{\mu}] \\
& \leq \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
& = 0, \\
& \frac{\partial g}{\partial q_{HH}}|_{q=\tilde{q}} = \frac{\lambda_{HH}}{\lambda_{LH}} \frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} \leq 0.
\end{aligned}$$

*Step 3:* Constructing a feasible solution.

Define  $(q^*, W^*, R^*)$  as follows:

$$\begin{aligned}
q^* &= \tilde{q}, \\
W_{ij}^* &= \begin{cases} \frac{q_{ij}^*}{\mu - \lambda_{HL} q_{HL}^* - \lambda_{LL} q_{LL}^*} & ij \in \{HL, LL\} \\ 0 & ij \in \{LH, HH\} \end{cases}, \\
R_{ij}^* &= \begin{cases} \Delta_v q_{LL}^* & ij = HL \\ 0 & ij \in \{LH, HH, LL\} \end{cases}.
\end{aligned}$$

By construction,  $(q^*, W^*, R^*)$  satisfies all IC and resource constraints. Furthermore, since  $Z(q^*, W^*, R^*) = g(\tilde{q})$ ,  $(q^*, W^*, R^*)$  is an optimal solution to the server's problem.  $\square$

**Proposition 10.** *The optimal mechanism is to fully admit the type-HL customers, partially admit the type-HH customers, and give the type-HH absolute priority, if the following conditions hold*

- $\Delta_v \geq \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH})^2}$ ,

- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL})^2} - \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{HH}}\right) \frac{1}{\mu} > 0,$
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH})^2} - \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{HH}}\right) \frac{\mu}{(\mu - \lambda_{HH})^2} \leq 0.$

*Proof. Step 1:* Constructing an upper bound function.

$$\begin{aligned}
& Z(q, W, R) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_c \lambda_{HH} W_{HH} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \lambda_{HL} \Delta_c W_{HH} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{HH}}\right) \frac{\lambda_{HH} q_{HH}}{\mu - \lambda_{HH} q_{HH}} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) \\
&\equiv g(q).
\end{aligned}$$

Here the first inequality is due to  $IC(HL - HH)$  and the nonnegativity of  $W$  and  $R$ , while the second one follows from  $RE(T)$  and  $RE(\{HH\})$ .

*Step 2:* Optimizing the upper bound function.

Define  $\tilde{q}_{HL} = 1$ ,  $\tilde{q}_{LH} = \tilde{q}_{LL} = 0$ , and  $\tilde{q}_{HH}$  to be the solution to

$$h(q_{HH}) \equiv v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH} q_{HH})^2} - \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) \frac{\mu}{(\mu - \lambda_{HH} q_{HH})^2} = 0.$$

We claim that  $\tilde{q}$  is the optimal solution to  $\max_q \{g(q) : 0 \leq q_{ij} \leq 1, \forall ij \in T\}$ .

By assumption, we have

$$\begin{aligned}
h(0) &= v_H - c_L \frac{\mu}{(\mu - \lambda_{HL})^2} - \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{HH}}\right) \frac{1}{\mu} > 0, \\
h(1) &= v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH})^2} - \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{HH}}\right) \frac{\mu}{(\mu - \lambda_{HH})^2} \leq 0.
\end{aligned}$$

Because  $h(q_{HH})$  is decreasing in  $q_{HH}$ , it must be true that  $0 < \tilde{q}_{HH} \leq 1$ .

Moreover, it follows that

$$\begin{aligned}
& \frac{\partial g}{\partial q_{HH}}|_{q=\tilde{q}} = \lambda_{HH}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})}] \\
& = \lambda_{HH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})}] \\
& = 0, \\
& \frac{\partial g}{\partial q_{HL}}|_{q=\tilde{q}} = \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2}] \\
& = \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{HH})^2}] \\
& > \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})}] \\
& = 0, \\
& \frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} = \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_c \frac{\mu}{(\mu - \lambda_{LH} \tilde{q}_{LH} - \lambda_{HH} \tilde{q}_{HH})^2}] \\
& = \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c] \\
& \leq \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH})^2}] \\
& \leq \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})^2}] \\
& = 0, \\
& \frac{\partial g}{\partial q_{LL}}|_{q=\tilde{q}} = \frac{\lambda_{LL}}{\lambda_{LH}} \frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} \leq 0.
\end{aligned}$$

Clearly,  $\tilde{q}$  satisfies the KKT conditions of the optimization problem  $\max_q \{g(q) : 0 \leq q_{ij} \leq 1, \forall ij \in T\}$ . Since  $g(q)$  is a concave function,  $\tilde{q}$  must be an optimal solution to this problem and  $g(\tilde{q})$  is an upper bound on  $Z(q, W, R)$ .

*Step 3:* Constructing a feasible solution.

Define  $(q^*, W^*, R^*)$  as follows:

$$\begin{aligned}
q^* &= \tilde{q}, \\
W_{ij}^* &= \begin{cases} \frac{q_{HH}^*}{\mu - \lambda_{HH} q_{HH}^*} & ij = HH \\ \frac{1}{\lambda_{HL}} \left( \frac{\lambda_{HH} q_{HH}^* + \lambda_{HL} q_{HL}^*}{\mu - \lambda_{HH} q_{HH}^* - \lambda_{HL} q_{HL}^*} - \frac{\lambda_{HH} q_{HH}^*}{\mu - \lambda_{HH} q_{HH}^*} \right) & ij = HL \\ 0 & ij \in \{LH, LL\} \end{cases}, \\
R_{ij}^* &= \begin{cases} \Delta_c W_{HH}^* & ij = HL \\ 0 & ij \in \{LH, HH, LL\} \end{cases}.
\end{aligned}$$

By construction,  $(q^*, W^*, R^*)$  satisfies all IC and resource constraints. Furthermore, since  $Z(q^*, W^*, R^*) = g(\tilde{q})$ ,  $(q^*, W^*, R^*)$  is an optimal solution to the server's problem.  $\square$

### Three Groups of Customers

**Proposition 11.** *The optimal mechanism is to fully admit the type-HL and type-LL customers, partially admit the type-HH customers, give the type-HH absolute priority, and pool the type-HL and type-LL customers in one queue, if the following conditions hold:*

- $\Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \leq \Delta_c \frac{1}{\mu}$ ,
- $\Delta_v(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} + \frac{\lambda_{HH} + \lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) \geq \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH})^2}$ ,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{HH}}) \frac{1}{\mu} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} > 0$ ,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} \leq 0$ .

*Proof. Step 1:* Constructing an upper bound function.

$$\begin{aligned}
& Z(q, W, R) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) \\
&\quad - \lambda_{HH} \Delta_v q_{LH} - \lambda_{LL}(\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH}) - \lambda_{HL}(\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH} + \Delta_v q_{LL}) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \lambda_{LL} q_{LL} - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \lambda_{HH} W_{HH} \\
&\quad + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} \lambda_{HH} q_{HH} - \Delta_v(1 + \frac{\lambda_{HH} + \lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) \lambda_{LH} q_{LH} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \lambda_{LL} q_{LL} - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\lambda_{HH} q_{HH}}{\mu - \lambda_{HH} q_{HH}} \\
&\quad + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} \lambda_{HH} q_{HH} - \Delta_v(1 + \frac{\lambda_{HH} + \lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) \lambda_{LH} q_{LH} \\
&\equiv g(q).
\end{aligned}$$

*Step 2:* Optimizing the upper bound function. Define  $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$ ,  $\tilde{q}_{LH} = 0$  and  $\tilde{q}_{HH}$  to

be the solution to

$$h(q_{HH}) \equiv v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH}q_{HH})^2} - \Delta_c \left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}\right) \frac{\mu}{(\mu - \lambda_{HH}q_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} = 0.$$

By assumption,

$$h(0) = v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_c \left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}\right) \frac{1}{\mu} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} > 0,$$

$$h(1) = v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH})^2} - \Delta_c \left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}\right) \frac{\mu}{(\mu - \lambda_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} \leq 0.$$

Because  $h(q_{HH})$  is decreasing in  $q_{HH}$ , it follows that  $0 < \tilde{q}_{HH} \leq 1$ . Next we verify that  $\tilde{q}$  satisfies the KKT conditions of the optimization problem  $\max_q \{g(q) : 0 \leq q_{ij} \leq 1, \forall ij \in T\}$ . Because  $g(q)$  is concave, these conditions are sufficient for optimality.

The partial derivatives at  $\tilde{q}$  satisfy

$$\begin{aligned}
& \frac{\partial g}{\partial q_{HH}}|_{q=\tilde{q}} = \lambda_{HH}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}] \\
& = \lambda_{HH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}] \\
& = 0,
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial g}{\partial q_{HL}}|_{q=\tilde{q}} = \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2}] \\
& = \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{LL} \tilde{q}_{HH})^2}] \\
& > \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}] \\
& = 0,
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial g}{\partial q_{LL}}|_{q=\tilde{q}} = \lambda_{LL}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
& = \lambda_{LL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
& > \lambda_{LL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}] \\
& = 0,
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} = \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_v(1 + \frac{\lambda_{HH} + \lambda_{LL} \lambda_{HL}}{\lambda_{LH}})] \\
& = \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_v(1 + \frac{\lambda_{HH} + \lambda_{LL} \lambda_{HL}}{\lambda_{LH}})] \\
& \leq \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} \tilde{q}_{HH})^2} - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH} \tilde{q}_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}] \\
& = 0.
\end{aligned}$$

*Step 3:* Constructing a feasible solution

Given  $\tilde{q}$ , we define a solution  $(q^*, W^*, R^*)$  as follows:

$$\begin{aligned}
q^* &= \tilde{q}, \\
W_{ij}^* &= \begin{cases} \frac{q_{HH}^*}{\mu - \lambda_{HH} q_{HH}^*} & ij = HH \\ \frac{1}{\lambda_{LL} + \lambda_{HL}} \left( \frac{\lambda_{HH} q_{HH}^* + \lambda_{LL} q_{LL}^* + \lambda_{HL} q_{HL}^*}{\mu - \lambda_{HH} q_{HH}^* - \lambda_{LL} q_{LL}^* - \lambda_{HL} q_{HL}^*} - \frac{\lambda_{HH} q_{HH}^*}{\mu - \lambda_{HH} q_{HH}^*} \right) & ij \in \{LL, HL\} \\ 0 & ij = LH, \end{cases} \\
R_{ij}^* &= \begin{cases} 0 & ij = LH \\ \Delta_v q_{LH}^* & ij = HH \\ \Delta_v q_{LH}^* + \Delta_c W_{HH}^* - \Delta_v q_{HH}^* & ij = HL \\ \Delta_v q_{LH}^* + \Delta_c W_{HH}^* - \Delta_v q_{HH}^* + \Delta_v q_{LL}^* & ij = LL. \end{cases}
\end{aligned}$$

By construction,  $(q^*, W^*, R^*)$  satisfies all IC and resource constraints. Furthermore, since  $Z(q^*, W^*, R^*) = g(\tilde{q})$ ,  $(q^*, W^*, R^*)$  is an optimal solution to the server's problem.  $\square$

**Proposition 12.** *If all of the following conditions holds, the optimal admission policy is to fully admit the type-HL and type-LL customers and partially admit the type-HH customers. The optimal priority ranking is absolute, with type-HH at the highest, followed by type-HL, and type-LL at the lowest. Strategic idleness is always required.*

- $\Delta_v \geq \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) \frac{\mu}{(\mu - \lambda_{HH})^2}$ ,
- $\Delta_v > \frac{\Delta_c}{\lambda_{LL}} \left( \frac{\lambda_{HH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} \right)$ ,
- $\Delta_v \left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} + \frac{\lambda_{HH} + \lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}\right) \leq \Delta_c \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} \frac{1}{\mu}$ ,
- $c_L < \Delta_c \frac{\lambda_{HH}}{\lambda_{LL} + \lambda_{HL}}$ ,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH})^2} - \Delta_c \left(1 + \frac{\lambda_{HL}}{\lambda_{HH}}\right) \frac{\mu}{(\mu - \lambda_{HH})^2} > 0$ .
- $v_L - c_L \frac{\Delta_c}{\Delta_v} > 0$ ,
- $v_H - \Delta_v \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) - c_L \frac{\Delta_c}{\Delta_v} \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) < 0$ .

*Proof. Step 1:* Constructing an upper bound function.

$$\begin{aligned}
& Z(q, W, R) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) \\
&\quad - \lambda_{LL} \Delta_c W_{LH} - \lambda_{HH} \max\{\Delta_v q_{LL} - \Delta_c W_{LL}, \Delta_v q_{LL} - \Delta_c W_{HL}, 0\} - \lambda_{HL} \max\{\Delta_v q_{LL}, \Delta_c W_{HH}\} \\
&\equiv g(q, W).
\end{aligned}$$

*Step 2:* Optimizing the upper bound function. Let  $(\tilde{q}, \tilde{W})$  be an optimal solution to  $\max_{q, W} \{g(q, W) : 0 \leq q_{ij} \leq 1, \forall ij \in T, (RE)\}$ . First of all, it can be verified that  $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$ , and  $\tilde{q}_{LH} = 0$  if  $\tilde{q}_{LL} < 1$ . And since  $W_{HH}$  doesn't appear in  $g(q, W)$ , we arbitrarily set it to its lower bound, i.e.  $\tilde{W}_{HH} = \frac{1}{\mu - \lambda_{HH}}$ . The problem left is to determine  $\tilde{q}_{LL}$ ,  $\tilde{W}_{LL}$  and  $\tilde{W}_{HL}$ .

Because  $c_L < \Delta_c \frac{\lambda_{HH}}{\lambda_{LL} + \lambda_{HL}}$ , it must be true that  $\max\{\Delta_v q_{LL} - \Delta_c W_{LL}, \Delta_v q_{LL} - \Delta_c W_{HL}, 0\} = 0$ , otherwise we can increase  $W_{LL}$  and  $W_{HL}$  equivalently to increase  $g(q, W)$ . It follows that

$$\Delta_v \tilde{q}_{LL} - \Delta_c \tilde{W}_{LL} \leq 0, \text{ and } \Delta_v \tilde{q}_{LL} - \Delta_c \tilde{W}_{HL} \leq 0.$$

This implies that strategic idleness has to be applied to the optimal solution, because otherwise even if  $LL$  is given the lowest priority, the left hand side of the first equality will always be positive, under the condition that  $\Delta_v > \frac{\Delta_c}{\lambda_{LL}} \left( \frac{\lambda_{HH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} \right)$ . Therefore, we have

$$\begin{aligned}
\tilde{W}_{LL} &= \frac{\Delta_v}{\Delta_c} \tilde{q}_{LL}, \\
\tilde{W}_{HL} &= \max\left\{ \frac{\Delta_v}{\Delta_c} \tilde{q}_{LL}, \frac{1}{\lambda_{HL}} \left( \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} - \frac{\lambda_{HH}}{\mu - \lambda_{HH}} \right) \right\},
\end{aligned}$$

since  $\tilde{W}_{HL}$  has to satisfy  $RE(\{HH, HL\})$  at the same time.

After substituting  $W_{LL}$  and  $W_{HL}$  out, it can be verified that  $g(q, W)$  has two breakpoint at  $q_{LL}^1 \equiv \frac{\Delta_c}{\Delta_v(\mu - \lambda_{HH})}$  and  $q_{LL}^2 \equiv \frac{\Delta_c}{\Delta_v \lambda_{HL}} \left( \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} - \frac{\lambda_{HH}}{\mu - \lambda_{HH}} \right)$ . Furthermore,  $0 < q_{LL}^1 < q_{LL}^2$ .

The marginal benefit of increasing  $q_{LL}$  is

$$\begin{aligned}
& v_L - c_L \frac{\Delta_c}{\Delta_v}, \quad 0 \leq q_{LL} \leq q_{LL}^1, \\
& v_H - \Delta_v \left( 1 + \frac{\lambda_{HL}}{\lambda_{LL}} \right) - c_L \frac{\Delta_c}{\Delta_v}, \quad q_{LL}^1 < q_{LL} \leq q_{LL}^2, \\
& v_H - \Delta_v \left( 1 + \frac{\lambda_{HL}}{\lambda_{LL}} \right) - c_L \frac{\Delta_c}{\Delta_v} \left( 1 + \frac{\lambda_{HL}}{\lambda_{LL}} \right), \quad q_{LL}^2 < q_{LL} \leq 1.
\end{aligned}$$

Because  $v_L - c_L \frac{\Delta_c}{\Delta_v} > 0$  and  $v_H - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) - c_L \frac{\Delta_c}{\Delta_v}(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) > 0$ , the optimal value of  $q_{LL}$  is either  $q_{LL}^1$  or  $q_{LL}^2$ , depending on the sign of  $v_H - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) - c_L \frac{\Delta_c}{\Delta_v}$ . Specifically, if  $v_H - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) - c_L \frac{\Delta_c}{\Delta_v} \leq 0$ ,  $(\tilde{q}, \tilde{W})$  is given below

$$\tilde{q}_{ij} = \begin{cases} 0 & ij = LH \\ 1 & ij \in \{HH, HL\} \\ \frac{\Delta_c}{\Delta_v(\mu - \lambda_{HH})} & ij = LL \end{cases},$$

$$\tilde{W}_{ij} = \begin{cases} 0 & ij = LH \\ \frac{1}{\mu - \lambda_{HH}} & ij = HH \\ \frac{1}{\lambda_{HL}} \left( \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} - \frac{\lambda_{HH}}{\mu - \lambda_{HH}} \right) & ij = HL \\ \frac{\Delta_v}{\Delta_c} \tilde{q}_{LL} & ij = LL \end{cases}.$$

If  $v_H - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) - c_L \frac{\Delta_c}{\Delta_v} > 0$ ,  $(\tilde{q}, \tilde{W})$  can be specified as below

$$\tilde{q}_{ij} = \begin{cases} 0 & ij = LH \\ 1 & ij \in \{HH, HL\} \\ \frac{\Delta_c}{\Delta_v \lambda_{HL}} \left( \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} - \frac{\lambda_{HH}}{\mu - \lambda_{HH}} \right) & ij = LL \end{cases}$$

$$\tilde{W}_{ij} = \begin{cases} 0 & ij = LH \\ \frac{1}{\mu - \lambda_{HH}} & ij = HH \\ \frac{\Delta_v}{\Delta_c} \tilde{q}_{LL} & ij = \{HL, LL\} \end{cases}.$$

*Step 3:* Constructing a feasible solution.

Given  $(\tilde{q}, \tilde{W})$ , we define a solution  $(q^*, W^*, R^*)$  as follows:

$$\begin{aligned} q^* &= \tilde{q}, \\ W^* &= \tilde{W}, \\ R_{ij}^* &= \begin{cases} 0 & ij \in \{LH, LL, HH\} \\ \max\{\Delta_v q_{LL}^*, \Delta_c W_{HH}^*\} & ij = HL \end{cases}. \end{aligned}$$

By construction,  $(q^*, W^*, R^*)$  satisfies all IC and resource constraints. Furthermore, since  $Z(q^*, W^*, R^*) = g(\tilde{q}, \tilde{W})$ ,  $(q^*, W^*, R^*)$  is an optimal solution to the server's problem.  $\square$

## Four Groups of Customers

**Proposition 13.** *If the following conditions hold, the optimal admission control is to fully admit the type-HL and type-LL customers and partially admit the type-HH and type-LH customers. If  $q_{HH}^* < 1$  and  $q_{LH}^* < 1$ , the optimal priority ranking is absolute, with type-HH at the highest, followed by type-LH at the second, and type-LL and type-HL equally at the lowest. If  $q_{HH}^* = 1$  and  $q_{LH}^* < 1$ , the optimal mechanism uses randomized ranking between type-HH and type-LH. If  $q_{HH}^* = 1$  and  $q_{LH}^* = 1$ , the optimal mechanism only uses two priority classes, with type-LH and type-HH equally at the higher, and type-LL and type-HL equally at the lower priority.*

- $\Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \leq \Delta_c \frac{1}{\mu}$ ,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{HH}}) \frac{1}{\mu} > 0$ ,
- $\Delta_v(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} + \frac{\lambda_{HH} + \lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) < \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{1}{\mu}$ ,

*Proof. Step 1:* Constructing an upper bound function.

$$\begin{aligned}
& Z(q, W, R) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) \\
&\quad - \lambda_{HH} \Delta_v q_{LH} - \lambda_{LL} \max\{\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH}, \Delta_c W_{LH}\} \\
&\quad - \lambda_{HL} \max\{\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH} + \Delta_v q_{LL}, \Delta_c W_{LH} + \Delta_v q_{LL}\} \\
&\equiv g(q, W).
\end{aligned}$$

*Step 2:* Optimizing the upper bound function.

Let  $(\tilde{q}, \tilde{W})$  be an optimal solution to  $\max_{q, W} \{g(q, W) : 0 \leq q_{ij} \leq 1, \forall ij \in T, \text{ (RE)}\}$ . Clearly,  $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$  in this case. The problem left is to find the optimal value of  $q_{HH}$ ,  $q_{LH}$ ,  $W_{HH}$  and  $W_{LH}$ . For fixed  $q$ , the optimal value of  $W$  depends on  $q_{HH}$ . If  $\tilde{q}_{HH} < 1$ , type-HH should be given absolute priority over LH, otherwise  $(\tilde{q}, \tilde{W})$  would be suboptimal since we could re-balance the allocations between type-LH and type-HH to improve  $g(q, W)$ . This implies

$$\tilde{W}_{HH} = \frac{\tilde{q}_{HH}}{\mu - \lambda_{HH} \tilde{q}_{HH}}.$$

Further more,  $RE(\{HH, LH\})$  should also be binding at optimality, which leads to

$$\tilde{W}_{LH} = \frac{1}{\lambda_{LH}} \left( \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\lambda_{HH}\tilde{q}_{HH}}{\mu - \lambda_{HH}\tilde{q}_{HH}} \right)$$

In addition, the following condition has to be satisfied at optimality in order minimize the information rent:

$$\Delta_v \tilde{q}_{LH} + \Delta_c \tilde{W}_{HH} - \Delta_v \tilde{q}_{HH} = \Delta_c \tilde{W}_{LH}.$$

First, we assume  $q_{HH} < 1$  and look for a solution that satisfies the first order conditions. Define  $q_{LL}^1 = q_{HL}^1 = 1$ , and  $(q_{HH}^1, q_{LH}^1)$  to be the solution of

$$v_H - c_L \frac{\mu}{(\mu - \lambda_{LL} - \lambda_{HL} - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH})^2} - \Delta_c \left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}\right) \frac{\mu}{(\mu - \lambda_{HH}q_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} = 0,$$

$$\begin{aligned} & v_H - c_L \frac{\mu}{(\mu - \lambda_{LL} - \lambda_{HL} - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH})^2} \\ & - \Delta_c \frac{\mu}{(\mu - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH})^2} - \Delta_v \left(1 + \frac{\lambda_{HH} + \lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}\right) = 0, \end{aligned}$$

$$\Delta_v q_{LH} + \Delta_c \frac{q_{HH}}{\mu - \lambda_{HH}q_{HH}} - \Delta_v q_{HH} = \Delta_c \frac{1}{\lambda_{LH}} \left( \frac{\lambda_{HH}q_{HH} + \lambda_{LH}q_{LH}}{\mu - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH}} - \frac{\lambda_{HH}q_{HH}}{\mu - \lambda_{HH}q_{HH}} \right),$$

where the left hands side of the first two equalities are the marginal benefits to increase  $q_{HH}$  and  $q_{LH}$  respectively.

If  $q_{HH}^1 \leq 1$ , we claim that  $(\tilde{q}, \tilde{W})$ , specified as below, is optimal to  $\max_{q,W} \{g(q, W) : 0 \leq q_{ij} \leq 1, \forall ij \in T, (RE)\}$ :

$$\begin{aligned} \tilde{q} &= q^1, \\ \tilde{W}_{ij} &= \begin{cases} \frac{\tilde{q}_{HH}}{\mu - \tilde{q}_{HH}} & ij = HH \\ \frac{1}{\lambda_{LH}} \left( \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\lambda_{HH}\tilde{q}_{HH}}{\mu - \lambda_{HH}\tilde{q}_{HH}} \right) & ij = LH \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} \left( \frac{\sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}} - \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}} \right) & ij = \{HL, LL\} \end{cases}. \end{aligned}$$

If  $q_{HH}^1 > 1$ , it follows that  $\tilde{q}_{HH} = 1$ . In order to minimize the information rent and maximize the social surplus, we need to apply randomized priority rule to reduce the difference in  $W_{HH}$  and  $W_{LL}$  such that the following equalities holds:

$$\Delta_v \tilde{q}_{LH} + \Delta_c \tilde{W}_{HH} - \Delta_v \tilde{q}_{HH} = \Delta_c \tilde{W}_{LH}.$$

The resource constraint  $RE(\{HH, LH\})$  should still be binding in this case. Therefore,

$$\lambda_{LH} \tilde{W}_{LH} + \lambda_{HH} \tilde{W}_{HH} = \frac{\lambda_{LH} \tilde{q}_{LH} + \lambda_{HH} \tilde{q}_{HH}}{\mu - \lambda_{LH} \tilde{q}_{LH} - \lambda_{HH} \tilde{q}_{HH}}.$$

Combining the above two equations yields

$$\tilde{W}_{LH} = \left( \frac{1}{\lambda_{LH} + \lambda_{HH}} \right) \left[ \frac{\lambda_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\Delta_v}{\Delta_c} \lambda_{HH} (1 - \tilde{q}_{LH}) \right].$$

To find the first order solution under the current setting, we define  $q_{LL}^2 = q_{HL}^2 = q_{HH}^2 = 1$  and  $q_{LH}^2$  to be the solution to the following equation

$$v_H - \Delta_v \left( 1 + \frac{\lambda_{HH}}{\lambda_{LH}} \right) - \Delta_c \left( 1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}} \right) \left( \frac{1}{\lambda_{LH} + \lambda_{HH}} \right) \left[ \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{LH}q_{LH})^2} + \frac{\Delta_v \lambda_{HH}}{\Delta_c} \right] = 0,$$

where the left hand side is the marginal benefit of increasing  $q_{LH}$  under this condition.

If  $q_{LL}^2 < 1$ ,  $(\tilde{q}, \tilde{W})$  can be specified as below:

$$\tilde{q} = q^2$$

$$\tilde{W}_{ij} = \begin{cases} \left( \frac{1}{\lambda_{LH} + \lambda_{HH}} \right) \left[ \frac{\lambda_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\Delta_v}{\Delta_c} \lambda_{HH} (1 - \tilde{q}_{LH}) \right] & ij = LH \\ \frac{1}{\lambda_{HH}} \left( \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \lambda_{LH}\tilde{W}_{LH} \right) & ij = HH \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} \left( \frac{\sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}} - \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}} \right) & ij \in \{HL, LL\} \end{cases}.$$

Finally, if  $q_{LL}^2 \geq 1$ , it is clear that  $\tilde{q}_{LH} = \tilde{q}_{HH} = \tilde{q}_{LL} = \tilde{q}_{HL} = 1$ , and  $\tilde{W}$  can be specified as follows:

$$\tilde{W}_{ij} = \begin{cases} \frac{1}{\mu - \lambda_{LH} + \lambda_{HH}} & ij = \{LH, HH\} \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} \left( \frac{\sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}} - \frac{\lambda_{HH} + \lambda_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}} \right) & ij = \{HL, LL\} \end{cases}.$$

*Step 3:* Constructing a feasible solution.

Given  $(\tilde{q}, \tilde{W})$ , we define a solution  $(q^*, W^*, R^*)$  as follows:

$$q^* = \tilde{q},$$

$$W^* = \tilde{W},$$

$$R^*_{ij} = \begin{cases} 0 & ij = LH \\ \Delta_v q^*_{LH} & ij = HH \\ \max\{\Delta_v q^*_{LH} + \Delta_c W^*_{HH} - \Delta_v q^*_{HH}, \Delta_c W^*_{LH}\} & ij = LL \\ R^*_{LL} + \Delta_v q^*_{LL} & ij = HL \end{cases}.$$

By construction,  $(q^*, W^*, R^*)$  satisfies all IC and resource constraints. Furthermore, since  $Z(q^*, W^*, R^*) = g(\tilde{q}, \tilde{W})$ ,  $(q^*, W^*, R^*)$  is an optimal solution to the server's problem.  $\square$

**Proposition 14.** *If the following conditions hold, the optimal admission control is to fully admit the type-HL and type-HH customers and partially admit the type-LL and type-LH customers. If  $q_{LL}^* < 1$  and  $q_{LH}^* < 1$ , the optimal priority ranking is absolute, with type-LH at the highest, followed by type-HH at the second highest, type-HL at the third, and type-LL at the lowest. If  $q_{HH}^* = 1$  and  $q_{LH}^* < 1$ , the optimal mechanism uses randomized ranking between LH and HH. If  $q_{HH}^* = 1$  and  $q_{LH}^* = 1$ , the optimal mechanism only uses two priority classes, with type-LH and type-HH equally at the higher priority, and type-LL and type-HL equally at the lower priority.*

- $\Delta_v \geq \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH})^2},$
- $\Delta_v \geq \frac{\Delta_c}{\lambda_{LL}} (\frac{\lambda_{HH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}}),$
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH})^2} > 0,$
- $c_L \geq \Delta_c \frac{\lambda_{HH} + \lambda_{HL}}{\lambda_{LL}},$
- $\Delta_v (\frac{\lambda_{HL}}{\lambda_{LL}} - \frac{\lambda_{HH}}{\lambda_{LH}}) > \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}).$

*Proof. Step 1:* Constructing an upper bound function.

$$\begin{aligned}
& Z(q, W, R) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) - \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) \\
&\quad - \lambda_{LL} \Delta_c W_{LH} - \lambda_{HH} \max\{\Delta_c W_{LH} + \Delta_v q_{LL} - \Delta_c W_{LL}, \Delta_v q_{LH}\} \\
&\quad - \lambda_{HL} \max\{\Delta_c W_{LH} + \Delta_v q_{LL} - \Delta_c W_{LL} + \Delta_c W_{HH}, \Delta_c W_{LH} + \Delta_v q_{LL}\} \\
&\equiv g(q, W).
\end{aligned}$$

*Step 2:* Optimizing the upper bound function.

Let  $(\tilde{q}, \tilde{W})$  be an optimal solution to  $\max_{q, W} \{g(q, W) : 0 \leq q_{ij} \leq 1, \forall ij \in T, \text{ (RE)}\}$ . It is obvious that  $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$ . So the problem left is to find the optimal  $q_{LL}, q_{LH}, W_{LH}, W_{HH}$  and  $W_{LL}$ .

We note that strategic idleness should never be used since the cost outweighs the benefit. Therefore, for fixed  $q$ , the optimal value for  $W_{LL}$  is equal to its upper bound, which equals to

$$\tilde{W}_{LL} = \frac{1}{\lambda_{LL}} \left( \frac{\sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}} - \frac{\lambda_{HH} \tilde{q}_{HH} + \lambda_{LH} \tilde{q}_{LH} + \lambda_{HL} \tilde{q}_{HL}}{\mu - \lambda_{HH} \tilde{q}_{HH} - \lambda_{LH} \tilde{q}_{LH} - \lambda_{HL} \tilde{q}_{HL}} \right).$$

The optimal value of  $W_{LH}$  and  $W_{HH}$  depends on  $q_{LL}$ . If  $\tilde{q}_{LL} < 1$ , it can be verified that  $LH$  should be given absolute priority over  $HH$ . This implies

$$\begin{aligned} \tilde{W}_{LH} &= \frac{\tilde{q}_{LH}}{\mu - \lambda_{LH} \tilde{q}_{LH}}, \\ \tilde{W}_{HH} &= \frac{1}{\lambda_{HH}} \left( \frac{\lambda_{HH} \tilde{q}_{HH} + \lambda_{LH} \tilde{q}_{LH}}{\mu - \lambda_{HH} \tilde{q}_{HH} - \lambda_{LH} \tilde{q}_{LH}} - \frac{\lambda_{LH} \tilde{q}_{LH}}{\mu - \lambda_{LH} \tilde{q}_{LH}} \right). \end{aligned}$$

If  $q_{LL} = 1$ , we may need to apply randomized priority ranking between  $LH$  and  $HH$ . In either case,  $(\tilde{q}, \tilde{W})$  should satisfy the following equality in order to minimize the information rent:

$$\Delta_c \tilde{W}_{LH} + \Delta_v \tilde{q}_{LL} = \Delta_v \tilde{q}_{LH} + \Delta_c \min\{\tilde{W}_{LL}, \tilde{W}_{HH}\}.$$

We note that  $g(q, W)$  has two breakpoints. The first breakpoint  $(q^a, W^a)$  occurs when  $W_{LL}^a \leq W_{HH}^a$ , while the second breakpoint  $(q^b, W^b)$  arises when  $q_{LL}^b = 1$ . Clearly  $q_{LL}^a < q_{LL}^b = 1$ , otherwise  $W_{LL}^a$  must be greater than  $W_{HH}^a$ . Our next step is to look for first order solutions separately on the three regimes  $0 \leq q_{LL} < q_{LL}^a$ ,  $q_{LL}^a \leq q_{LL} < 1$ , and  $q_{LL} = 1$

For the first regime  $0 \leq q_{LL} \leq q_{LL}^a$ , let  $(q^1, W^1)$  be the first order solution. Clearly,  $q_{HH}^1 = q_{HL}^1 = 1$ , and  $(q_{LL}^1, q_{LH}^1)$  is the solution of

$$\begin{aligned} v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL} q_{LL} - \lambda_{LH} q_{LH})^2} - \Delta_v \left(1 + \frac{\lambda_{HL}}{\lambda_{LL}}\right) &= 0, \\ v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL} q_{LL} - \lambda_{LH} q_{LH})^2} - \Delta_v \left(1 + \frac{\lambda_{HH}}{\lambda_{LH}}\right) - \Delta_c \left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}\right) \frac{\mu}{(\mu - \lambda_{LH} q_{LH})^2} &= 0, \\ \Delta_c \frac{q_{LH}}{\mu - \lambda_{LH} q_{LH}} + \Delta_v q_{LL} &= \Delta_v q_{LL} + \Delta_c \frac{1}{\lambda_{LL}} \left( \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \frac{\lambda_{HH} q_{HH} + \lambda_{LH} q_{LH} + \lambda_{HL} q_{HL}}{\mu - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH} - \lambda_{HL} q_{HL}} \right), \end{aligned}$$

where the left hand sides of the first two equalities are the marginal benefits of increasing  $q_{LL}$  and  $q_{LH}$  respectively.

In this case,  $W^1$  is determined by

$$W_{ij}^1 = \begin{cases} \frac{q_{LH}^1}{\mu - q_{LH}^1} & ij = LH \\ \frac{1}{\lambda_{HH}} \left( \frac{\lambda_{HH} q_{HH}^1 + \lambda_{LH} q_{LH}^1}{\mu - \lambda_{HH} q_{HH}^1 - \lambda_{LH} q_{LH}^1} - \frac{\lambda_{HH} q_{LH}^1}{\mu - \lambda_{HH} q_{LH}^1} \right) & ij = HH \\ \frac{1}{\lambda_{HL}} \left( \frac{\lambda_{HL} q_{HL}^1 + \lambda_{HH} q_{HH}^1 + \lambda_{LH} q_{LH}^1}{\mu - \lambda_{HL} q_{HL}^1 - \lambda_{HH} q_{HH}^1 - \lambda_{LH} q_{LH}^1} - \frac{\lambda_{HH} q_{HH}^1 + \lambda_{LH} q_{LH}^1}{\mu - \lambda_{HH} q_{HH}^1 - \lambda_{LH} q_{LH}^1} \right) & ij = HL \\ \frac{1}{\lambda_{LL}} \left( \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}^1}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}^1} - \frac{\lambda_{HL} q_{HL}^1 + \lambda_{HH} q_{HH}^1 + \lambda_{LH} q_{LH}^1}{\mu - \lambda_{HL} q_{HL}^1 - \lambda_{HH} q_{HH}^1 - \lambda_{LH} q_{LH}^1} \right) & ij = LL \end{cases}.$$

If  $W_{LL}^1 \leq W_{HH}^1$ , it follows that  $q_{LL}^1 \leq q_{LL}^a$ . Because  $g(q, W)$  is concave,  $(\tilde{q}, \tilde{W}) = (q^1, W^1)$  is optimal. If  $W_{LL}^1 > W_{HH}^1$ , we need to continue searching in the second regime  $q_{LL}^a < q_{LL} \leq q_{LL}^b$ . Define  $q_{HH}^2 = q_{HL}^2 = 1$ , and  $(q_{LL}^2, q_{LH}^2)$  to be the solution of

$$v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}q_{LL} - \lambda_{LH}q_{LH})^2} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) = 0,$$

$$v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}q_{LL} - \lambda_{LH}q_{LH})^2} - \Delta_v(1 + \frac{\lambda_{HH}}{\lambda_{LH}}) - \Delta_c(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) \frac{\mu}{(\mu - \lambda_{LH}q_{LH})^2} = 0,$$

$$\Delta_c \frac{q_{LH}}{\mu - \lambda_{LH}q_{LH}} + \Delta_v q_{LL} = \Delta_v q_{LL} + \Delta_c \frac{1}{\lambda_{HH}} \left( \frac{\lambda_{HH}q_{HH} + \lambda_{LH}q_{LH}}{\mu - \lambda_{HH}q_{HH} - \lambda_{LH}q_{LH}} - \frac{\lambda_{LH}q_{LH}}{\mu - \lambda_{LH}q_{LH}} \right).$$

Again, the left hand sides of the first two equalities are the marginal benefits of increasing  $q_{LL}$  and  $q_{LH}$  under the current assumptions.

If  $q_{LL}^2 \leq 1$ , the optimal solution  $(\tilde{q}, \tilde{W})$  can be specified as below:

$$\tilde{q} = q^2,$$

$$\tilde{W}_{ij} = \begin{cases} \frac{q_{LH}^2}{\mu - q_{LH}^2} & ij = LH \\ \frac{1}{\lambda_{HH}} \left( \frac{\lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2} - \frac{\lambda_{HH}q_{LH}^2}{\mu - \lambda_{HH}q_{LH}^2} \right) & ij = HH \\ \frac{1}{\lambda_{HL}} \left( \frac{\lambda_{HL}q_{HL}^2 + \lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HL}q_{HL}^2 - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2} - \frac{\lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2} \right) & ij = HL \\ \frac{1}{\lambda_{LL}} \left( \frac{\sum_{ij \in T} \lambda_{ij}q_{ij}^2}{\mu - \sum_{ij \in T} \lambda_{ij}q_{ij}^2} - \frac{\lambda_{HL}q_{HL}^2 + \lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HL}q_{HL}^2 - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2} \right) & ij = LL \end{cases}.$$

However, if  $q_{LL}^2 > 1$ , we need to look for the optimal solution in the third regime  $q_{LL} = 1$ . Let  $(q^3, W^3)$  be the first order condition in this case. Apparently,  $q_{LL}^3 = q_{HL}^3 = q_{HH}^3 = 1$ . The problem left is to determine the optimal  $q_{LH}$ ,  $W_{LH}$  and  $W_{HH}$ .

First, we fix  $q_{LH}$  and determine the optimal  $W_{LH}$  and  $W_{HH}$ . Recall that the optimal solution satisfies

$$\Delta_c W_{LH} + \Delta_v q_{LL} = \Delta_v q_{LH} + \Delta_c W_{HH}.$$

Additionally,  $RE(\{HH, LH\})$  is binding at optimality, implying

$$\lambda_{LH}W_{LH} + \lambda_{HH}W_{HH} = \frac{\lambda_{LH}q_{LH} + \lambda_{HH}q_{HH}}{\mu - \lambda_{LH}q_{LH} - \lambda_{HH}q_{HH}}.$$

Combining the above two equations yields

$$W_{LH} = \left( \frac{1}{\lambda_{LH} + \lambda_{HH}} \right) \left[ \frac{\lambda_{HH} + \lambda_{LH}q_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}q_{LH}} - \frac{\Delta_v}{\Delta_c} \lambda_{HH}(1 - q_{LH}) \right].$$

Under this setting,  $q_{LH}^3$  should be the solution of the following equation

$$v_H - \Delta_v \left(1 + \frac{\lambda_{HH}}{\lambda_{LH}}\right) - \Delta_c \left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}\right) \left(\frac{1}{\lambda_{LH} + \lambda_{HH}}\right) \left[\frac{\mu}{(\mu - \lambda_{HH} - \lambda_{LH}q_{LH})^2} + \frac{\Delta_v \lambda_{HH}}{\Delta_c}\right] = 0,$$

where the left hand side is the marginal benefit of increasing  $q_{LH}$  in this case.

If  $q_{LL}^3 < 1$ ,  $(\tilde{q}, \tilde{W})$  can be specified as below:

$$\begin{aligned} \tilde{q} &= q^3, \\ \tilde{W}_{ij} &= \begin{cases} \left(\frac{1}{\lambda_{LH} + \lambda_{HH}}\right) \left[\frac{\lambda_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\Delta_v}{\Delta_c} \lambda_{HH} (1 - \tilde{q}_{LH})\right] & ij = LH \\ \frac{1}{\lambda_{HH}} \left(\frac{\lambda_{HH}\tilde{H}\tilde{H} + \lambda_{LH}\tilde{L}\tilde{H}}{\mu - \lambda_{HH}\tilde{H}\tilde{H} - \lambda_{LH}\tilde{L}\tilde{H}} - \lambda_{LH}\tilde{W}_{LH}\right) & ij = HH \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} \left(\frac{\sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}} - \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}}\right) & ij = \{HL, LL\} \end{cases}. \end{aligned}$$

Finally, if  $q_{LL}^3 \geq 1$ , it is obvious that  $\tilde{q}_{LH} = \tilde{q}_{HH} = \tilde{q}_{LL} = \tilde{q}_{HL} = 1$ , and  $\tilde{W}$  satisfies

$$\tilde{W}_{ij} = \begin{cases} \frac{1}{\mu - \lambda_{LH} + \lambda_{HH}} & ij = \{LH, HH\} \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} \left(\frac{\sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij}\tilde{q}_{ij}} - \frac{\lambda_{HH} + \lambda_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}}\right) & ij = \{HL, LL\} \end{cases}.$$

*Step 3:* Constructing a feasible solution.

Given  $(\tilde{q}, \tilde{W})$ , we define a solution  $(q^*, W^*, R^*)$  as follows:

$$\begin{aligned} q^* &= \tilde{q}, \\ W^* &= \tilde{W}, \\ R_{ij}^* &= \begin{cases} 0 & ij = LH \\ \Delta_c W_{LH}^* & ij = LL \\ \Delta_v q_{LH}^* & ij = HH \\ \Delta_c W_{LH}^* + \Delta_v q_{LL}^* & ij = HL \end{cases}. \end{aligned}$$

By construction,  $(q^*, W^*, R^*)$  satisfies all IC and resource constraints. Furthermore, since  $Z(q^*, W^*, R^*) = g(\tilde{q}, \tilde{W})$ ,  $(q^*, W^*, R^*)$  is an optimal solution to the server's problem.  $\square$