Designing Service Menus for Bipartite Queueing Systems

René Caldentey[†] Varun Gupta[†] Lisa Hillas[†]

Abstract

We consider a multi-class multi-server queueing system, in which customers of different types have heterogeneous preferences over the many servers available. The goal of the service provider is to design a menu of service classes that balances two competing objectives: (1) maximize customers' average matching reward and (2) minimize customers' average waiting time. A service class corresponds to a single queue served by a subset of servers under a FCFS-ALIS service discipline. Customers act as rational self-interested utility maximizing agents when choosing which service class to join. In particular, they join the class that maximizes their expected ex-ante net utility, which is given by the difference between the server-dependent service reward they receive minus a disutility based on the mean steady-state waiting time of the service class they join. We study the problem under (conventional) heavy traffic conditions, that is, in the limit as the traffic intensity of the system approaches one from below. For the case of two servers, we provide a complete and insightful characterization of the possible menus and their delay-reward tradeoffs. For general number of servers, we prove that if the service provider only cares about minimizing average delay or maximizing total matching reward then very simple menus are optimal. Finally, we provide Mixed Integer Linear Programming (MILP) formulations for optimizing the delay-reward trade-off within a fairly rich and practically relevant families of menus, which we term *Partitioned* and *Tailored.*

Keywords: Multi-class queueing system; first-come-first-served; bipartite matching; steady-state analysis.

1 Introduction

This paper is concerned with the problem of designing a queueing matching system in a multi-class and multi-server service environment, in which customers of different types arrive to the system seeking service by one of many available servers. Servers are heterogeneous in terms of the amount of time it takes them to serve a customer as well as on other attributes that affect the reward that customers receive for the service. The goal of the service provider is to design a service mechanism that will match customers to servers and will balance two (usually) competing objectives: (1) maximize customers' average service reward and (2) minimize customers' average waiting time. We will restrict ourselves to a special class of mechanisms in which the service provider offers a static menu of service classes and customers choose which one of them to join upon arrival. A service class is defined by a single queue served by a specific subset of servers under a FCFS-ALIS service discipline[†]. We will assume

[†] Booth School of Business, The University of Chicago. Email: {rene.caldentey,varun.gupta,lhillas}@chicagobooth.edu

[†] The acronym FCFS-ALIS stands for "first come first served - assign longest idle server". This means that when a server becomes idle it selects the customer who has been waiting the longest among those that it can serve. Similarly, a customer that can be served by multiple idle servers selects the server that has been idle the longest.

that customers act as rational self-interested utility maximizers when choosing which service class to join. In particular, they join the class that maximizes their expected net utility, which is given by the difference between the server-dependent service reward they receive minus a disutility waiting cost based on the mean steady-state waiting time of the service class they join.

To illustrate some of the features of the problem at hand, let us consider a concrete example with two servers. In this setting, the service provider can offer one of the five different service menus in Figure 1. For example, she can offer a *Dedicated menu* (far-most left panel) consisting of two service classes (queues) each served exclusively by one of the two servers. Alternatively, the service provider can offer a *Full menu* (middle panel) in which customers have three options; they can choose between two dedicated service classes each served exclusively by one of the two servers or they can join a third class served by both servers. A customer who chooses this third class does not know with certainty which one of the two servers will be the one providing the service.

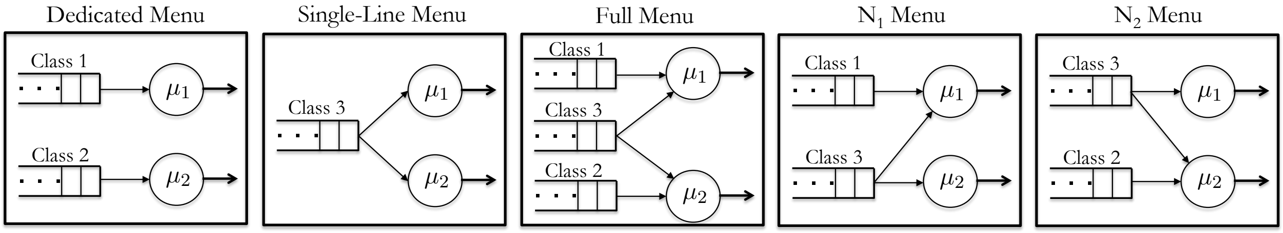

Figure 1: Possible service menus in a system with two servers.

Figure 2 depicts an example of the equilibrium performance of the five menus in Figure 1 in the average reward vs. average delay quadrant for different values of the system utilization ρ. A complete analysis of the two-server case is presented in Section 4.

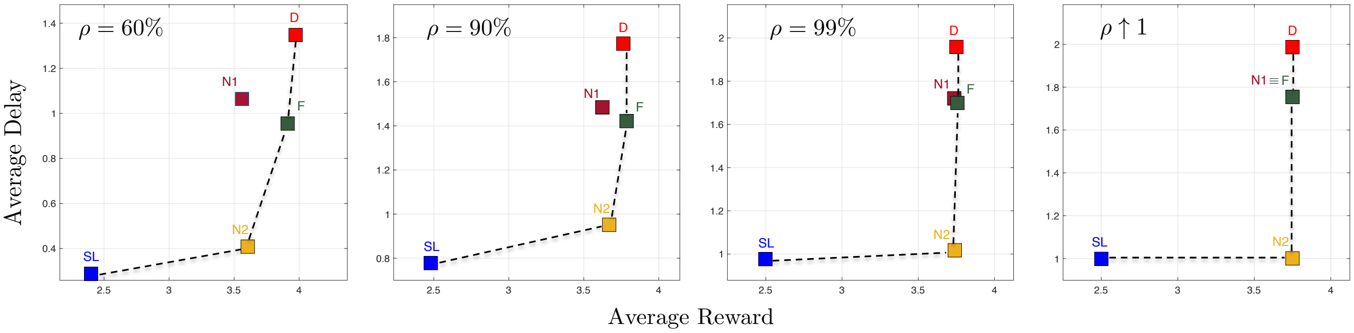

Figure 2: Equilibrium performance of the five menus (Dedicated (D), Single Line (SL), Full (F), N_1 and N_2) in the average reward vs. average delay quadrant for different values of the system utilization ρ.

For example, we will show that if the service provider only cares about minimizing customers' average waiting times, then the *Single-Line menu* is an optimal menu (see Theorem 3), which is something to be expected since a single line guarantees complete resource pooling. On the flip side, if the service provider is exclusively interested in maximizing average matching rewards and pays no attention to waiting times, then the *Dedicated menu* is an optimal menu (see Theorem 4) under heavy traffic conditions.

It is worth noticing that in this example, menu N_2 dominates the other four menus when the system operates in heavy traffic $\rho \uparrow 1$ (right-most panel). Thus, in this case it is possible to select a menu that achieves first best performance simultaneously on both measures (see Remark 4 for details). Interestingly, since the N_2 menu is a restricted version of the Full menu in which service class 1 is not offered, this example shows that reducing customers' choices can lead to more efficient outcomes (a form of Braess's paradox).

Related Literature. The specific class of queueing systems that we consider roots back to the early work of Kaplan (1984, 1988) who proposed a multi-class multi-server queueing system operating under a FCFS service discipline to model public housing waiting lists and to determine steady-state matching assignments, i.e., the probability that an applicant (families and individuals) of a certain type ends up receiving a particular class of public housing. (See also the early studies by Schwartz (2004), Green (1985) and the recent paper by Castro et al. (2020) on the 2-server N model under a FCFS discipline, and the paper by Talreja and Whitt (2008) for more general matching topologies.) The queueing formulation in Kaplan (1988) was later adapted by Caldentey and Kaplan (2002) who introduced an *infinite bipartite matching* model to investigate the problem of determining matching probabilities under a FCFS service discipline. The infinite matching model was further developed in Caldentey et al. (2009), Bušić et al. (2013), Adan and Weiss (2012), Adan et al. (2018a) and Fazel-Zarandi and Kaplan (2018). The connection between the steady-state solution of the queueing model and the infinite bipartite matching model was formalized by Adan and Weiss (2014) under the FCFS-ALIS service discipline (see also Adan et al., 2018b, 2019 and the survey by Gardner and Righter, 2020).

For the most part, the aforementioned stream of literature has assumed that the *matching topology* connecting services classes to servers is exogenously given and has focused on the performance analysis of the queueing system; i.e., identifying conditions that ensure stability or characterizing steady-state matching rates. The problem of designing optimal matching topologies is studied in Afèche et al. (2021) under the assumption that consumers are passive agents who do not choose which service class to join. In this setting, they can restrict themselves to topologies in which there is a one-to-one correspondence between customer types and service classes and so the design problem reduces to deciding the subset of servers that should serve each customer class. To deal with the combinatorial complexity of the problem identified by Adan and Weiss (2014), Afèche et al. (2021) rely on a heavy traffic analysis that unveils a surprisingly simple structure. Namely, under heavy-traffic conditions, they show that any bipartite matching system can be partitioned into a collection of complete resource pooling (CRP) subsystems, which are interconnected by means of a direct acyclic graph (DAG). The significance of these results is that they allow to replace the combinatorial structure of the original queueing system (expressed in terms of permutations of servers) by a more aggregate representation defined by the collection of topological orders of the CRP components. As result, they show that the DAG together with the aggregate service capacity on each CRP component fully determine the vector of steady-state waiting times. Combining this insight together with a Quadratic Program approach to approximate matching flows, Afèche et al. (2021) propose a mixed-integer linear program formulation that can be used to characterize the set of matching topologies that optimize the tradeoff between matching rewards and waiting times in a Pareto efficiency sense.

Our paper builds on and extends Afèche et al. (2021) by allowing consumers to choose the service class they want to join. As we will show in the following sections, this generalization is not trivial.

For one, the number of service classes can no longer be reduced to the number of customer types as the service provider can in principle offer a full service menu with as many service classes as the number of possible subsets of the servers. Also, by allowing customers to self-select the service class they want to join, the service provider has less control over the final matching. In other words, while in Afèche et al. (2021) the service provider acts as a *central planner* that has full control on how to route customers to service classes, in our case the central planner acts as a *principal* that can only induce *agents* (customers) to join a particular service class by designing an *incentive compatible* menu. The Principal-Agent nature of our problem implies that waiting times and matching flows must be computed imposing equilibrium conditions, which brings an extra layer of complexity to the problem. Finally, another subtle but important difference between Afèche et al. (2021) and our paper relates to how a heavy traffic analysis can be conducted. Specifically, in Afèche et al. (2021) the heavy traffic limit was essentially exogenously defined by letting the vector of customers' arrival rates converge (from below) along a pre-specified direction to a limiting vector of arrival rates. In contrast, in our case in which customers self-select the service class they want to join, the direction of convergence to heavy traffic is endogenously determined in equilibrium.

A distinctive feature of the papers that we have discussed so far, and which is also central to our work, is the FCFS-ALIS service discipline that is used in the matching of customers and servers. This type of service discipline is appropriate in settings (such as public housing allocations, adoption agencies or state-run nursing homes, to name a few) in which fairness considerations and/or legal regulations prevent the service provider from implementing other type of priority-based policies that could be (or could be perceived to be) discriminatory. If we relax this requirement, there exists a vast queueing literature on skill-based routing devoted to the problem of characterizing dynamic scheduling policies to control and optimize the flow of customers in a multi-server setting. Some representative examples of this stream of work include Harrison (1998), Harrison and Lopez (1999), Mandelbaum and Stolyar (2004), Atar (2005), Bell and Williams (2005), Wallace and Whitt (2005), Dai and Tezcan (2005), Gurvich and Whitt (2009, 2010), and Ward and Armony (2013).

Another stream of papers that is relevant to our work is concerned with the design of differentiated service menus. Some representative papers in this area include Van Mieghem (2000), Plambeck (2004), Maglaras and Zeevi (2005), Afèche (2013), Afèche and Pavlin (2016), Nazerzadeh and Randhawa (2018), Afèche et al. (2021), Ashlagi et al. (2021) and Ashlagi et al. (2022). The typical setting in these paper is one in which customers are heterogeneous in terms of their valuation or willingness-to-pay for service and their sensitivity to delay, while servers are homogeneous (in many cases a single server is considered). Under these conditions, a service class consists of two components: (1) the price that the service provider charges for the service and (2) the expected waiting time. Operationally, the service provider controls the service discipline which allows her to offer differentiated waiting times to the different service classes. The goal of the service provider is to design a menu of service classes that maximizes her profit or in some cases a social welfare objective.

In terms of applications, stochastic matching systems have been extensively used in the healthcare literature to study organ transplantation (e.g., Zenios et al. 2000, Akan et al. 2012, Bertsimas et al. 2013 and Ding et al. 2018) and kidney exchanges (e.g., Unver (2010), Anderson et al. 2017, Ashlagi et al. 2019 and Akbarpour et al. 2018). Other applications include public housing (e.g., Bloch and Cantala 2017, Leshno 2017, and Arnosti and Shi 2018), adoptions (e.g., Baccara et al. 2014 and Slaugh et al. 2016), labor markets (e.g., Rogerson et al. 2005, Arnosti et al. 2018 and Baccara et al. 2020), assemble-to-order manufacturing (e.g., Gurvich and Ward, 2014 and Nazari and Stolyar, 2019) and process flexibility (e.g., Jordan and Graves, 1995, Bassamboo et al., 2012, Tsitsiklis and Xu, 2012, 2017 and Shi et al., 2019).

2 Model Description

In this section we provide a detailed mathematical description of the model and basic definitions. Figure 3 provides a schematic illustration of the queueing system and the main notation.

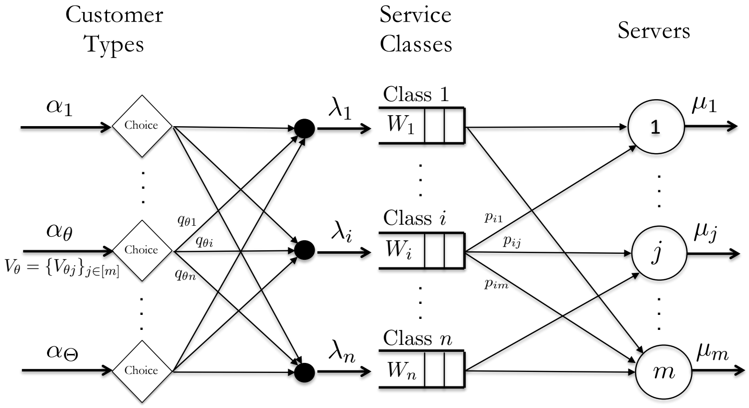

Figure 3: A multi-class multi-server matching queueing system.

A collection Θ of customer types arrives to the system over time according to independent Poisson processes with rates $\alpha = \{\alpha_\theta\}_{\theta \in [\Theta]}$. (For a positive integer k, we let $[k] := \{1, 2, \ldots, k\}$). The service system is composed of m parallel servers with exponentially distributed service times with rate $\mu = \{\mu_j\}_{j \in [m]}$. Customers have heterogeneous preferences over the servers and we denote by $V_\theta = \{V_{\theta j}\}_{j \in [m]}$ the vector of rewards for a type-θ customer, where $V_{\theta j}$ is the reward that a customer θ gets when served by server j.

In an effort to maximize the quality of the matching between customers and servers, the service provider offers n service classes, where each service class $i \in [n]$ is defined by a subset $S_i \subseteq [m]$ of servers that can serve those customers joining class i. Such a collection can be expressed by a binary compatibility matrix $M \in \{0, 1\}^{n \times m}$ where the entries of M specify which service classes can be served by which servers. That is, server class i can be served by server j if and only if $m_{ij} = 1$ for $i \in [n]$ and $j \in [m]$. We assume that upon arrival and before observing the queue lengths, customers select one service class and join the queue of this class and wait to be served according to a FIFO queueing discipline. This decision is irreversible, that is, after joining a queue, the customer stays in it until the service is completed. Servers, on the other hand, serve the different classes using a FCFS service discipline. We also assume that an arriving customer who finds an empty queue will be routed to the compatible server that has been idle the longest. Under these conditions, we say that the queueing system operates under a FCFS-ALIS (first come first served - assign longest idle server) service discipline.

A *strategy* for an arriving customer type θ is a probability distribution $q_\theta = \{q_{\theta i}\}_{i \in [n]}$ over the set of service classes, where $q_{\theta i}$ is the probability that a type-θ customer selects to join class $i \in [n]$. We let $q = \{q_\theta\}_{\theta \in \Theta}$ denote the strategy profile of all customer types and let $\mathcal{Q}$ be the set of all feasible strategy profiles, that is,

$$\mathcal{Q} := \Big\{ q \in \mathbb{R}_+^{\Theta \times n} \colon \sum_{i \in [n]} q_{\theta i} = 1 \text{ for all } \theta \in [\Theta] \Big\}.$$

A strategy profile $q \in \mathcal{Q}$ induces a vector of arrival rates $\lambda(q) = \{\lambda_i(q)\}_{i \in [n]}$ to each service class, where

$$\lambda_i(q) = \sum_{\theta \in [\Theta]} \alpha_\theta \, q_{\theta i} \qquad i = 1, \ldots, n.$$

Indirectly, through the vector of arrival rates $\lambda(q)$, the strategy profile q also determines the vector $W(q) = \{W_i(q)\}_{i \in [n]}$ of steady-state waiting times for each class as well as the matrix of matching probabilities $p(q) = (p_{ij}(q) \colon i \in [n], j \in [m])$ between service classes and servers, where $p_{ij}(q)$ denotes the steady-state probability that a customer joining class i will be served by server j under the strategy profile q. We will restrict our attention to menus M and strategies q that are stable in the sense that they jointly admit a well-defined steady state for the service system.

Proposition 1. (Adan and Weiss, 2014, Theorem 2.1) *The menu M and strategy q admit a steady state under a FCFS-ALIS service discipline if and only if the following condition is satisfied:*

$$\sum_{j \in \mathscr{S}} \mu_j > \sum_{i \in U_{\mathscr{S}}(M)} \lambda_i(q) \qquad \textit{for all } \mathscr{S} \subseteq [m],$$

where $U_{\mathscr{S}}(M)$ is the subset of service classes that can only be served by servers in $\mathscr{S}$.

We will further restrict attention to *admissible* menus M for which the stability condition above is satisfied for some feasible strategy profile q.

Definition 1. (Admissible Menus and Strategies) *A menu M is* admissible *if there exists a strategy profile $q = \{q_\theta\}_{\theta \in \Theta}$ for which the service system is stable, that is, the inequalities in Proposition 1 are satisfied. We let $\mathcal{M}$ denote the set of admissible menus.*

For an admissible menu $M \in \mathcal{M}$, we denote by $\mathcal{Q}(M) \subseteq \mathcal{Q}$ the set of all feasible strategy profiles for which the service system is stable.

A necessary and sufficient condition for $\mathcal{M}$ to be non-empty is that the cumulative arrival rate is strictly less than the cumulative service capacity, $|\alpha| < |\mu|$. (For a vector $x = (x_i)_{i=1}^k$, we let $|x| := \sum_{i=1}^k x_i$). Indeed, under this condition it is not hard to see that any menu M in which every server is connected to at least one service class is admissible. In particular, the single-line menu (i.e., $n = 1$ and $m_{1j} = 1$ for all $j \in [m]$) is admissible.

We assume that the expected utility that a type-θ customer gets from joining class i is equal to

$$U_{\theta i}(W, p) := \sum_{j \in S_i} p_{ij} \, V_{\theta j} - \delta \, W_i,$$

where δ is a scalar parameter capturing customers' sensitivity to delays. Given a pair (W, p) of steady-steady waiting times and matching probabilities, a rational utility-maximizing type-θ customer joins the service class i that maximizes $U_{\theta i}(W, p)$ in equilibrium.

Definition 2. (ϵ-Equilibrium and Equilibrium Profiles) *Let $M \in \mathcal{M}$ and let $q^* = \{q^*_\theta\}_{\theta \in [\Theta]} \in \mathcal{Q}$ be a strategy profile with corresponding vector of waiting times $W^* = \{W_i(q^*)\}_{i \in [n]}$ and matrix of matching probabilities $p^* = [p_{ij}(q^*)]_{i \in [n], j \in [m]}$.*

–) ϵ-Equilibrium Profile*: For a given $\epsilon \geq 0$, we say that (q^*, W^*, p^*) is an* ϵ-equilibrium *profile if for all $\theta \in [\Theta]$ and for all $i, k \in [n]$*

$$q^*_{\theta i}\Big(U_{\theta i}(W^*, p^*) - U_{\theta k}(W^*, p^*)\Big) + \epsilon \geq 0.$$

We let $\mathcal{Q}^\epsilon(M)$ be the set of strategies q^* for which an ϵ-equilibrium profile (q^*, W^*, p^*) exists.

–) Equilibrium Profile*: We say that (q^*, W^*, p^*) is an* equilibrium *profile if it is a 0-equilibrium profile. We let $\mathcal{Q}^*(M)$ be the set of strategies q^* for which an equilibrium profile (q^*, W^*, p^*) exists.*

Trivially, every equilibrium profile is an ϵ-equilibrium profile for all $\epsilon > 0$. The following theorem guarantees the existence of equilibrium profiles when the system has sufficient service capacity to serve all of the customers.

Theorem 1. *Suppose that $|\alpha| < |\mu|$, and $M \in \mathcal{M}$ is an admissible service menu. Then there exists an equilibrium strategy profile $q^* \in \mathcal{Q}(M)$.*

The proof of this and other results can be found in Appendix A.

The final component of the model corresponds to the objective that the service provider uses to select an optimal menu M^*. Similar to the preferences of individual customers, we assume that the service provider is interested in maximizing the value generated by the matching between customers and servers while minimizing the waiting time experienced by these customers. Specifically, for a given admissible menu M and consumers' strategy q, we assume that the service provider collects a payoff equal to

$$\Pi(M, q) := \overline{V}(M, q) - \zeta\, \overline{W}(M, q), \tag{1}$$

where ζ is a positive scalar that captures the service provider's sensitivity to customers' delays and

$$\overline{V}(M, q) := \sum_{\theta \in [\Theta]} \sum_{i \in [n]} \sum_{j \in [m]} \alpha_\theta\, q_{\theta i}\, p_{ij}(q) V_{\theta j} \qquad \text{and} \qquad \overline{W}(M, q) := \sum_{i \in [n]} \lambda_i(q)\, W_i(q)$$

correspond to the cumulative steady state matching reward and waiting time, respectively, experienced by all consumers.

It is worth noticing that while the service provider is able to select the service menu M, it is the consumers who decide which service classes they want to join by selecting an equilibrium strategy $q^* \in \mathcal{Q}^*(M)$. Hence, the service provider's optimization problem can be formulated as follows:

$$\sup_{M \in \mathcal{M}} \ \sup_{q^* \in \mathcal{Q}^*(M)} \ \Pi(M, q^*). \tag{2}$$

Remark 1. Formulation (2) assumes that the service provider is able to select which equilibrium strategy $q^* \in \mathcal{Q}^*(M)$ consumers' will end up playing. This is, of course, without loss of generality for those admissible menus M for which $\mathcal{Q}^*(M)$ is a singleton. However, when M induces multiple equilibria formulation (2) models the problem of an 'optimistic' service provider. Alternatively, we could have adopted a pessimistic view by formulating the service provider's problem as follows:

$$\sup_{M \in \mathcal{M}} \ \inf_{q^* \in \mathcal{Q}^*(M)} \Pi(M, q^*). \ \diamond$$

Remark 2. (Social Planner) If $\zeta = \delta$ then $\Pi(M,q) = \sum_{\theta\in[\Theta]} \alpha_\theta \, \overline{U}_\theta(q)$, that is, service provider acts as a social planner who is interested in maximizing the cumulative utility of all consumers. $\diamond$

2.1 Roadmap of Analysis and Results

Before moving into the analysis of the service provider's problem, let us provide a summary of the methodology that we have used to tackle the problem of designing an optimal menu of service classes and the type of results that we have been able to obtain.

One of the major challenges in solving the optimization problem in (2) is the underlying combinatorial structure of the steady-state distribution of the system, which makes it difficult to calculate waiting times and matching rates beyond relatively small instances with $\max\{n,m\} \leq 12$ (see Adan and Weiss, 2014 and Afèche et al., 2021 for further discussion). To circumvent this challenge we will study the problem under (conventional) heavy traffic conditions, that is, in the limit as the traffic intensity of the system approaches one from below. In Section 3, we lay out the details of this heavy traffic regime and show how to compute mean waiting times and matching probabilities. A distinctive operating characteristic of the queueing system under heavy traffic conditions is that servers and service classes are partitioned into complete resource pooling components (CRP), which are interconnected by means of a directed acyclic graph (DAG). Furthermore, we show in Section 3.2 that under some mild conditions any limiting vector of waiting times can be *implemented* using a chained DAG configuration. This observation has important implications as it drastically simplifies the problem of characterizing an optimal service menu.

In Section 4, we illustrate the use of heavy traffic analysis to design an optimal service menu for a system with only two servers ($m = 2$). The purpose of this section is to highlight some key features of the problem in a setting in which we can provide a complete characterization of an optimal menu as a function of the model's parameters. In particular, the solution to the two-server case reveals that the different possible service menus can be partitioned into two main groups: (a) menus that achieve the minimum possible average waiting time and (b) menus that achieve the maximum possible average matching reward. Intuitively, delay-minimizing menus are those that are able to induce *complete capacity pooling* while reward-maximizing menus are those able to implement the matching that a central planner would select if she had complete control over the assignment of customers to servers. Interestingly, with two servers, every admissible service menu falls in one of these two categories. The solution also shows that out of the set of delay-minimizing menus the single-line produces the lowest possible average matching reward while out of the reward-maximizing menus the dedicated menu (i.e., one service class per server) generates the longest waiting-time delays. In other words, these two simple and commonly used menus are complete opposite designs when it comes to balancing the trade-off between average matching reward and mean customers' delays. Further, both of these menus are Pareto dominated by other service menus with more complex matching topologies. The single-line menu is dominated by a menu that achieves strictly higher average reward while preserving *complete capacity pooling*. On the other hand, the dedicated menu is dominated by another menu that results from *chaining* the dedicated service classes.

In Section 5, we investigate conditions under which a first best menu exists in a system with an arbitrary number of servers. We show that in the extreme cases in which the service provider's sensitivity to delay ζ is either zero or infinity an optimal menu is given by a Single-Line or Dedicated menu, respectively.

For an arbitrary value of ζ, we derive necessary (Theorem 6) and sufficient (Theorem 7) conditions for a first best outcome to be achieved, which are based on the solution to a max-flow problem.

In Section 6, we study a special class of *Partition menus* in which the set of servers are partitioned into pools of servers, each acting as a 'super-server' that serves a single service class. One of the key advantages of partition menus is that they are very simple to explain and implement in practice. Furthermore, despite their limitations, partition menus have a number of desirable theoretical properties (e.g., they include delay-minimizing or reward-maximizing menus) that the service provider can use to balance the trade-off between waiting times and matching values. Furthermore, they are also tractable from a computational standpoint and we exploit this in Section 6.3 to propose a mixed-integer linear program (MILP) that finds an optimal partition menu.

In Section 7 we adopt a *mechanism design* approach to tackle the problem of finding optimal service menus. Specifically, we interpret the service provider's problem as one in which she wants to design a different service class for each customer type, i.e., a menu with $n = |\Theta|$ in which every customer type joins a different (and unique) service class in equilibrium. We call this class of menus *Tailored menu* as every service class is tailored to a specific customer type. In Section 7.1 we develop a MILP formulation that finds a value-maximize menu among the class of tailored menus that support complete resource pooling and have minimum waiting times. In Section 7.2 we take the opposite point of view and formulate a MILP that finds a delay-minimizing tailored menu among those that generate maximum matching value.

Finally, in Section 8 we conduct a set of numerical experiments to compare the performance of Partition and Tailored menus as a function of different parameters of the model including the matrix of matching rewards V and the service provider's sensitivity to delay ζ.

3 Heavy Traffic Regime

In this section, we present the model that we will use to formally study the question of menu design through heavy-traffic asymptotics. First, in Section 3.1, we present the specific heavy traffic scaling of the system primitives. Then, in Section 3.2, we discuss how to calculate steady-state waiting times in heavy traffic and also recap a number of formulas derived in Afèche et al. (2021) and Caldentey et al. (2022) for this purpose. In Section 3.3 we present a quadratic programming (QP) formulation that we will use to approximate the matching probabilities under the FCFS-ALIS service requirement. Finally, in Section 3.4, we introduce the notion of a *heavy traffic* equilibria, which we use to extend Definition 2 to our heavy traffic regime.

3.1 Scaling

We construct a sequence of matching queueing systems parameterized by ϵ and use the superscript (ϵ) to emphasize the dependence of various quantities on ϵ. For example, $\alpha_\theta{}^{(\epsilon)}$ and $q_\theta{}^{(\epsilon)} = (q_{\theta 1}{}^{(\epsilon)}, \ldots, q_{\theta n}{}^{(\epsilon)})$ are the arrival rate and strategy profile of type-θ customers in system ϵ.

We assume that the bipartite matching system approaches heavy traffic as $\epsilon \downarrow 0$. Specifically, we assume that there are two vectors $A, a \in \mathbb{R}_+^\Theta$ (independent of ϵ) with $|A| = |\mu|$ so that the sequence of arrival rates $\alpha^{(\epsilon)} = \{\alpha_\theta{}^{(\epsilon)}\}_{\theta\in\Theta}$ satisfies (for ϵ small enough):

$$\alpha_\theta{}^{(\epsilon)} = A_\theta - a_\theta \epsilon \geq 0 \quad \text{for all } \theta \in [\Theta]. \tag{3}$$

Intuitively, in the heavy-traffic regime, the arrival rates $\alpha^{(\epsilon)}$ approach the limiting rates A along the direction specified by a. It follows that the traffic intensity of the ϵ^{th} system equals

$$\rho^{(\epsilon)} := \frac{\sum_\theta \alpha_\theta{}^{(\epsilon)}}{\mu_1 + \cdots + \mu_m} = \frac{|A| - |a|\,\epsilon}{|\mu|} = 1 - \frac{|a|}{|\mu|}\,\epsilon$$

and approaches one (i.e., 100% system utilization) as $\epsilon \downarrow 0$.

We let $\mathcal{M}^{(\epsilon)}$ denote the class of menus M that are admissible in the sense of Definition 1. It is not hard to see that the sets $\mathcal{M}^{(\epsilon)}$ are monotonic in ϵ and so the limit $\widehat{\mathcal{M}} := \lim_{\epsilon\downarrow 0} \mathcal{M}^{(\epsilon)}$ exists. We will refer to $\widehat{\mathcal{M}}$ as the set of *admissible menus in heavy traffic*.

Under the heavy traffic condition in (3), the waiting time $W_i{}^{(\epsilon)}(q^{(\epsilon)})$ will grow out of bound as $\epsilon \downarrow 0$. For this reason, we assume that $\delta^{(\epsilon)}$ goes to 0 as $\epsilon \downarrow 0$ in such a way that the product $\delta^{(\epsilon)}\, W_i{}^{(\epsilon)}(q^{(\epsilon)})$ converges to a well-defined non-trivial limit. In particular, we will assume that $\delta^{(\epsilon)} = \delta\,\epsilon$ for some fixed constant $\delta > 0$ independent of ϵ[†]. Given this scaling, we find convenient to define the scaled mean waiting time

$$\widehat{W}_i{}^{(\epsilon)}(q^{(\epsilon)}) := \epsilon \cdot W_i{}^{(\epsilon)}(q^{(\epsilon)}), \tag{4}$$

which remains bounded in heavy traffic. Finally, the expected utility of a customer type θ under strategy $q_\theta^{(\epsilon)}$ is given by

$$U_{\theta i}^{(\epsilon)}(q^{(\epsilon)}) = \sum_{j\in S_i} p_{ij}^{(\epsilon)}(q^{(\epsilon)})\, V_{\theta j} - \delta\, \widehat{W}_i{}^{(\epsilon)}(q^{(\epsilon)}).$$

Note that the valuations $V = [V_{\theta j}]$ and service rates $\mu = (\mu_j)$ remain constant independent of ϵ.

3.2 Mean Waiting Time in Heavy Traffic

To study equilibria under heavy traffic conditions, we need to be able to compute limiting scaled mean waiting times $\lim_{\epsilon\downarrow 0} \widehat{W}^{(\epsilon)}(q^{(\epsilon)})$ for a given sequence pre-limit strategy profiles $\{q^{(\epsilon)}\}_{\epsilon>0}$. In this section we present a high level summary of the results in Caldentey et al. (2022), who provide a detailed study on how to derive steady-state waiting times under heavy traffic conditions.

Let us fix an admissible menu $M \in \widehat{\mathcal{M}}$ in heavy traffic and let us consider a sequence of feasible strategy profiles $q^{(\epsilon)} \in \mathcal{Q}(M)$ for all $\epsilon > 0$. Motivated by our characterization of heavy traffic equilibria in Section 3.4, we consider strategy profiles that converge along a specific direction. Specifically, we assume that there exists $\hat{q} \in \mathcal{Q}$ and $\hat{\phi} \in \mathbb{R}^{|\Theta|\times n}$ such that $q^{(\epsilon)} = \hat{q} + \epsilon\,\hat{\phi} \in \mathcal{Q}$ for all $\epsilon \geq 0$. These strategy profiles induce a vector $\lambda^{(\epsilon)}$ of pre-limit arrival rates into service classes given by

$$\lambda_i{}^{(\epsilon)} = \sum_{\theta\in\Theta} \alpha_\theta{}^{(\epsilon)} q_{\theta i}{}^{(\epsilon)} = \sum_{\theta\in\Theta} A_\theta \hat{q}_{\theta i} - \epsilon \sum_{\theta\in\Theta} (a_\theta \hat{q}_{\theta i} - A_\theta \hat{\phi}_{\theta i}) + o(\epsilon) =: \Lambda_i - \epsilon\gamma_i + o(\epsilon). \tag{5}$$

In an effort to simplify the exposition, in what follows we assume that the limiting strategy profile $\hat{q}$ is such that $\Lambda_i > 0$ for all $i \in [n]$. The general case is discussed in Caldentey et al. (2022).

[†] Alternatively, we could consider a slightly more general scaling of $\delta^{(\epsilon)}$ that only requires $\lim_{\epsilon\downarrow 0} \frac{\delta^{(\epsilon)}}{\epsilon} = \delta$.

Given the menu M and the vectors of service rates $\mu = \{\mu_j\}_{j\in[m]}$ and arrival rates $\Lambda = (\Lambda_i \colon i \in [n])$, we define the corresponding *residual matching* $\breve{M} = [\breve{m}_{ij}] \in \{0,1\}^{n\times m}$ by removing those edges in M that must have zero flow in the limit as $\epsilon \downarrow 0$. To be precise, $\breve{m}_{\ell k} = 1$ if and only if $m_{\ell k} = 1$ and there exists a non-negative flow $f = [f_{ij}]$ such that

$$\sum_{i\in[n]} m_{ij}\, f_{ij} = \mu_j, \quad \forall j \in [m]; \quad \sum_{j\in[m]} m_{ij}\, f_{ij} = \Lambda_i, \quad \forall i \in [n]; \quad \text{and} \quad f_{\ell k} > 0.$$

Recall that the heavy traffic scaling in (3) implies that $|\Lambda| = |\mu|$.

Intuitively, for a service class ℓ and a server k with $m_{\ell k} = 1$ but $\breve{m}_{\ell k} = 0$, the flow of customers from ℓ to k must vanish in the heavy-traffic limit. The significance of the residual matching $\breve{M}$ is that its connected components induce a partition of the service classes and servers into a collection of *complete resource pooling (CRP)* components that establish a hierarchy of how congestion and delays build among the different service classes in the system.

Definition 3. (CRP Component) *Given the tuple (n, m, Λ, μ, M) and the induced residual matching $\breve{M}$, we say that the subset $\mathbb{C} = (\mathcal{C}, \mathcal{S}) \in 2^{[n]} \times 2^{[m]}$ of service classes and servers forms a CRP component if for any pair of nodes $k_1, k_2 \in \mathcal{C} \cup \mathcal{S}$ there exists a path between k_1 and k_2 in $\breve{M}$, and $\mathbb{C}$ is maximal in the sense that the condition is violated for any strict superset of $\mathbb{C}$.*

Intuitively, the "well-connectedness" within a CRP component allows the shifting of load from one service class to another on short time scales, and in particular under FCFS-ALIS policy to balance the waiting times in such a way that service classes that belong to the same CRP component have the same limiting scaled mean waiting time in heavy traffic (see Theorem 2).

We let $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_K\}$ denote the collection of CRP components induced by the residual matching $\breve{M}$, where K is the number of components and each $\mathbb{C}_k = (\mathcal{C}_k, \mathcal{S}_k)$ is defined by the subset $\mathcal{C}_k$ of service classes and the subset $\mathcal{S}_k$ of servers that belong to $\mathbb{C}_k$. Also, for each CRP component $\mathbb{C}_k$, we must have $\sum_{j\in\mathcal{S}_k} \mu_j = \sum_{i\in\mathcal{C}_k} \Lambda_i$, and we let $\widetilde{\gamma}_k := \sum_{i\in\mathcal{C}_k} \gamma_i$ denote its scaled capacity slack. From (3) we have that $|\widetilde{\gamma}| = |a|$.

The matching M induces a unique directed acyclic graph (DAG) on the CRP components of M in the following manner: The DAG includes an arc from CRP component $\mathbb{C}_{k_1}$ to CRP component $\mathbb{C}_{k_2}$ if, and only if, there exists a service class $i \in \mathcal{C}_{k_1}$ and a server $j \in \mathcal{S}_{k_2}$ such that $m_{ij} = 1$. Each arc in this DAG[‡] indicates that the destination CRP can absorb some of the load of, and must be more congested than, the origin CRP. The DAG therefore reflects the partial order on the congestion of the CRP components that emerges in heavy traffic.

In order to formally establish the connection between the DAG on the CRP components and the mean waiting times of the different service classes we need to introduce one additional element, namely, the collection of *topological orders* that the DAG induces over the CRP components. The importance of these topological orders comes from the fact that the state-space representation for the FCFS-ALIS matching model involves ranking the busy servers based on the order of the waiting time of the customers they are serving (see Adan and Weiss 2014 for details). As was proved in Afèche et al. (2021), in heavy-traffic this entails restricting attention to only certain permutations of the CRP components which have asymptotically non-zero steady-state probability. These permutations are precisely the topological orders of the DAG.

[‡] The claim that the resulting direct graph defined in this manner is in fact acyclic is formally proven in (Afèche et al., 2021, Lemma 2).

Definition 4. (Topological Orders of CRP Components) *Consider a matching $M \in \widehat{\mathcal{M}}$ and let $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_K\}$ be the collection of all CRP components of M. We say that a permutation $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(K))$ of $[K]$ induces a* topological order $(\mathbb{C}_{\sigma(1)}, \mathbb{C}_{\sigma(2)}, \ldots, \mathbb{C}_{\sigma(K)})$ *of the CRP components of M if for every directed arc $(\mathbb{C}_i, \mathbb{C}_j)$ from component $\mathbb{C}_i$ to component $\mathbb{C}_j$ in the DAG associated to M, we have $\sigma^{-1}(j) < \sigma^{-1}(i)$. We denote by $\mathcal{T}(M)$ the set of all permutations σ that induce a topological order and by $T(M)$ the cardinality of $\mathcal{T}(M)$.*

Following Afèche et al. (2021), for a given topological order $\sigma \in \mathcal{T}(M)$, let us define the unnormalized probability $\mathbb{Q}(\sigma)$ and the conditional waiting time $w_{\sigma,k}$ of CRP component $\mathbb{C}_k$ as follows:

$$\mathbb{Q}(\sigma) := \prod_{\kappa=1}^{K} \frac{1}{\sum_{\ell=1}^{\kappa} \widetilde{\gamma}_{\sigma(\ell)}} \qquad \text{and} \qquad w_{\sigma,k} := \sum_{\kappa=\sigma^{-1}(k)}^{K} \frac{1}{\sum_{\ell=1}^{\kappa} \widetilde{\gamma}_{\sigma(\ell)}}. \tag{6}$$

The following theorem is proven in Caldentey et al. (2022), which expresses the limiting scaled mean waiting times, $\widehat{W}_i^*$ in Definition 7 in terms of the values of $\mathbb{Q}(\sigma)$ and $w_{\sigma,k}$.

Theorem 2. *(Caldentey et al., 2022) For a given admissible service menu $M \in \widehat{\mathcal{M}}$ and a strategy profile $\hat{q} + \epsilon\hat{\phi}$ such that $\Lambda > 0$ in (5), let $\breve{M}$ be the residual matching and $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_K\}$ be the collection of CRP components induced by $\breve{M}$. Then, service classes that belong to the same CRP component experience the same scaled steady-state mean waiting time in heavy traffic. Furthermore, the scaled steady-state mean waiting time of CRP component $\mathbb{C}_k$ is equal to*

$$\widehat{W}_{\mathbb{C}_k} = \sum_{t=1}^{T(M)} \left(\frac{\mathbb{Q}(\sigma_t)}{\mathbb{Q}(\sigma_1) + \mathbb{Q}(\sigma_2) + \cdots + \mathbb{Q}(\sigma_{T(M)})} \right) w_{\sigma_t,k}. \tag{7}$$

The following is an immediate corollary of Theorem 2, which provide conditions under which complete resource pooling is possible.

Corollary 1. *Under the same conditions as in Theorem 2, $\widehat{W}_{\mathbb{C}_k} \geq 1/|a|$ for all $k \in [K]$. Furthermore, $\widehat{W}_{\mathbb{C}_{\hat{\kappa}}} = 1/|a|$ for some $\hat{\kappa} \in [K]$ if and only if there exists a directed path from $\mathbb{C}_{\hat{\kappa}}$ to any other CRP component $\mathbb{C}_k$ with $k \in [K] \setminus \{\hat{\kappa}\}$. This condition is trivially satisfied if the system has a single CRP component (i.e., $K = 1$).*

It is worth noticing that according to Theorem 2 the only information that is needed to compute the scaled steady-state mean waiting times in heavy traffic is the aggregated structure of the matching M in terms of CRP components and its DAG and topological orders together with the vector of scaled capacity slack $\widetilde{\gamma} = (\widetilde{\gamma}_1, \widetilde{\gamma}_2, \ldots, \widetilde{\gamma}_K)$. The more granular information about the specific compatibility structure between service classes and servers or the average arrival rates Λ and service capacities μ do not affect the computations of the waiting times. This is a key observation for the purpose of deriving optimal service menus as it simplifies the representation of the vector of limiting scaled waiting times that can be implemented. This motivates the following definition.

Definition 5. (Implementable Waiting Times) *Given a collection $\mathbb{C} = \{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_K\}$ of CRP components and a cumulative capacity slack $|a| > 0$, we say that a vector of limiting scaled waiting times $W = (W_1, W_2, \ldots, W_K)$ is* implementable *if there exist a DAG in $\mathbb{C}$ and a vector of scaled capacity slacks $\widetilde{\gamma} = (\widetilde{\gamma}_1, \widetilde{\gamma}_2, \ldots, \widetilde{\gamma}_K)$ with $|\widetilde{\gamma}| = |a|$ such that W_k is equal to $\widehat{W}_{\mathbb{C}_k}$ in (7) for all $k \in [K]$.*

In general, characterizing the set of implementable waiting times is challenging given the underlying combinatorial structure in (7). Proposition 2 below characterizes a special class which will prove useful later in our derivation of optimal service menus. The distinguishing feature of this class of waiting times is that they can be implemented using a *chained* DAG.

Definition 6. (Chained DAGs) *A DAG on* $\mathbb{C} = \{\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_K\}$ *is* chained *if there exists a partition* $\{\mathscr{C}_1, \mathscr{C}_1, \dots, \mathscr{C}_L\}$ *of* $\mathbb{C}$ *such that the permutation* $\sigma = (\sigma(1), \sigma(2), \dots, \sigma(K))$ *induced by any of its topological orders satisfies:* $\sigma^{-1}(j) < \sigma^{-1}(i)$ *if and only if there exist* $\ell_1, \ell_2 \in [L]$ *with* $\ell_1 \leq \ell_2$ *such that* $\mathbb{C}_i \in \mathscr{C}_{\ell_1}$ *and* $\mathbb{C}_j \in \mathscr{C}_{\ell_2}$*.*

Figure 4 illustrates two examples of a chained DAG over a collection of nine CRP components. For panel (a) on the left panel, $L = 6$ and $\mathscr{C}_1 = \{\mathbb{C}_2\}$, $\mathscr{C}_2 = \{\mathbb{C}_4\}$, $\mathscr{C}_3 = \{\mathbb{C}_1, \mathbb{C}_6, \mathbb{C}_7\}$, $\mathscr{C}_4 = \{\mathbb{C}_5, \mathbb{C}_8\}$, $\mathscr{C}_5 = \{\mathbb{C}_9\}$ and $\mathscr{C}_6 = \{\mathbb{C}_3\}$.

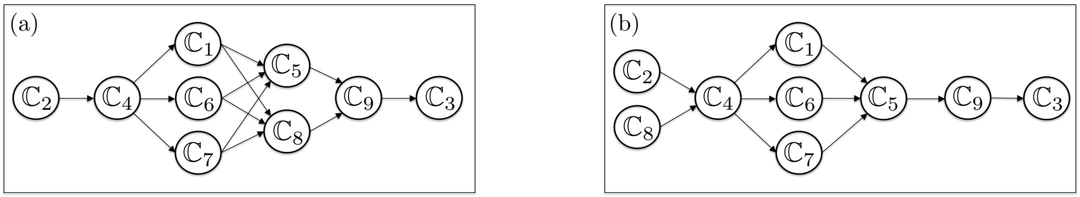

Figure 4: Two examples of chained DAGs over nine CRP components.

In the statement of the proposition, we donate by $(W_{(1)}, W_{(2)}, \dots, W_{(K)})$ the vector of order statistics of $(W_1, W_2, \dots, W_K)$.

Proposition 2. *Suppose* $W = (W_1, W_2, \dots, W_K)$ *satisfies* $1/|a| = W_{(1)} < W_{(2)}$*. Then,* W *is implementable by a chained DAG.*

From Corollary 1, we can interpret the condition in Proposition 2 as requiring W to be the limiting scaled waiting times of a DAG containing a CRP component that achieves complete resource pooling.

3.3 Matching Probabilities under Heavy Traffic Equilibrium

While the problem of computing waiting times under a FCFS-ALIS service discipline simplifies significantly under heavy traffic conditions, computing the matching probabilities p_{ij} remains computational challenging due to the underlying combinatorial structure of the state-space of the system (see Adan and Weiss, 2014). Caldentey et al. (2009) and Afèche et al. (2021) identify a special class of topologies (including spanning forests and complete and quasi-complete graphs) under which the matching probabilities can be computed efficiently. Specifically, for a given menu $M = [m_{ij}]$ in this class of topologies, the matching probabilities p_{ij} can be computed solving the following quadratic program:

$$\begin{aligned} \min_{p} \quad & \sum_{i\in[n]} \sum_{j\in[m]} \frac{\lambda_i}{\mu_j}\, m_{ij}\, p_{ij}^2 \qquad \textbf{(QP)} \\ \text{subject to} \quad & \sum_{i\in[n]} \lambda_i\, m_{ij}\, p_{ij} = \mu_j \quad \forall j \in [m], \\ & \sum_{j\in S_i} m_{ij}\, p_{ij} = 1 \quad \forall i \in [m], \\ & p_{ij} \geq 0 \quad \forall (i,j) \in [n]\times[m]. \end{aligned}$$

In general, however, the **QP** formulation only provides an approximation to the actual FCFS-ALIS matching probabilities (see Afèche et al., 2021 and Fazel-Zarandi and Kaplan, 2018 for detailed numerical experiments) and, to the best of our knowledge, it is still unknown whether there exists a computationally efficient method to determine the exact matching probabilities for an arbitrary matching topology.

In what follows, we will proceed with our analysis using the QP formulation to compute the matching flows. The following facts about an optimal solution to (**QP**) are proven in Afèche et al. (2021).

Proposition 3. *Let M be an admissible menu in heavy traffic, i.e., $M \in \widehat{\mathcal{M}}$. Then,*

1. *The quadratic program in (**QP**) is feasible and admits a unique optimal solution $p^*(M)$.*
2. *A feasible matrix of matching probabilities $p(M) = [p_{ij}(M)]$ is the optimal solution to (**QP**) if and only if there exist multipliers $(\omega_j,\ j \in [m])$ for the first set of constraints and $(\eta_i,\ i \in [n])$ for the second set of constraints satisfying the KKT first order stationarity conditions:*

$$p^*_{ij}(M) = \max\{\mu_j\,(\eta_i + \omega_j)\,, 0\}, \quad \forall (i,j) : m_{ij} = 1.$$

The second property is particularly useful because it allows for a simple encoding of the constraints imposed by the QP formulation[§] on the matching probabilities.

3.4 Heavy-Traffic Equilibrium

For a given admissible menu in heavy traffic $M \in \widehat{\mathcal{M}}$, we are interested in identifying a limiting equilibrium, as $\epsilon \downarrow 0$. To this end, we introduce the notion of a *heavy-traffic* equilibrium.

Definition 7. (Heavy Traffic Equilibrium) *For a given admissible menu in heavy traffic $M \in \widehat{\mathcal{M}}$, we say that $(\hat{q}^*, \widehat{W}^*, \hat{p}^*)$ is a* heavy traffic equilibrium *if there exists a vector $\hat{\phi}^* \in \mathbb{R}^{|\Theta|\times n}$ such that $\hat{q}^* + \epsilon\,\hat{\phi}^* \in \mathcal{Q}$ for all $\epsilon \geq 0$ and the following two conditions are satisfied:*

(a) Heavy Traffic Limit*: $\widehat{W}^* = \lim_{\epsilon\downarrow 0} \widehat{W}^{(\epsilon)}(\hat{q}^* + \epsilon\,\hat{\phi}^*)$ and $\hat{p}^* = \lim_{\epsilon\downarrow 0} p^{(\epsilon)}(\hat{q}^* + \epsilon\,\hat{\phi}^*)$.*

(b) Best-Response*: For all $\theta \in \Theta$ and for all $i, k \in [n]$*

$$\hat{q}^*_{\theta i}\left(U_{\theta i}(\widehat{W}^*, \hat{p}^*) - U_{\theta k}(\widehat{W}^*, \hat{p}^*)\right) \geq 0.$$

[§] Which is an approximation for the FCFS-ALIS requirements that we need to impose on $p = [p_{ij}]$.

We let $\widehat{\mathcal{Q}}^*(M)$ be the set of all strategy profiles $\hat{q}^*$ for which there exists a heavy traffic equilibrium $(\hat{q}^*, \widehat{W}^*, \hat{p}^*)$.

For the notion of a heavy-traffic equilibrium to be of any practical interest, we would like to be able to map it back to some concrete equilibrium in the pre-limit. The following proposition formalizes this requirement by showing that we can always view a heavy-traffic equilibrium as the limit of a sequence of ϵ-equilibria in the pre-limit, as $\epsilon \downarrow 0$.

Proposition 4. *Let $\hat{q}^* \in \widehat{\mathcal{Q}}^*(M)$ for some admissible menu $M \in \widehat{\mathcal{M}}$ in heavy traffic. Then, there exists a sequence of strategy profiles $(q^{(\epsilon)})_{\epsilon>0}$ with corresponding steady-state waiting times $W^{(\epsilon)} = \{W_i^{(\epsilon)}(q^{(\epsilon)})\}_{i\in[n]}$ and matching probabilities $p^{(\epsilon)} = [p_{ij}^{(\epsilon)}(q^{(\epsilon)})]_{i\in[n],j\in[m]}$ such that $(q^{(\epsilon)}, W^{(\epsilon)}, p^{(\epsilon)})$ is a $\Delta^{(\epsilon)}$-equilibrium profile for a sequence $(\Delta^{(\epsilon)})_{\epsilon>0}$ that satisfies $\lim_{\epsilon\downarrow 0} \Delta^{(\epsilon)} = 0$.*

Remark 3. A possible shortcoming of the definition of a heavy traffic equilibrium in Definition 7 is that the sequence of strategy profiles $\{q^{(\epsilon)}\}_{\epsilon>0}$ that defines a heavy traffic equilibrium is not required to be a sequence of pre-limit equilibria. Thus, it is possible that a heavy traffic equilibrium is not the limit of any sequence of pre-limit equilibria. Proposition 4, however, guarantees that the strategy profiles $\{q^{(\epsilon)}\}_{\epsilon>0}$ are ϵ-equilibria in the pre-limit and so the incentives that customers have to deviate from the strategy $q^{(\epsilon)}$ become negligible as $\epsilon \downarrow 0$. $\diamond$

The definition of a heavy-traffic equilibrium highlights an important feature of our asymptotic analysis of an equilibrium. Namely, to characterize a heavy traffic equilibrium it is not enough to identify the limiting strategy $\hat{q}^*$ but we must also specify the direction $\hat{\phi}^*$ of convergence. The reason is that the limiting vector of steady-state waiting times $\widehat{W}^*$ is not just a function of $\hat{q}^*$but also of $\hat{\phi}^*$. We illustrate this point with the following example.

Example. *Consider the system in Figure 5 with two customer types ($|\Theta| = 2$), two servers ($m = 2$) and two service classes ($n = 2$) each served exclusively by one of the servers. The arrival and service rates in the ϵ^{th} system are given by $\alpha^{(\epsilon)} = A - a\,\epsilon = (2,1) - (1,0)\,\epsilon$ and $\mu = (\mu_1, \mu_2) = (1,2)$, respectively. Customers of type 1 prefer server 1 over server 2 (i.e., $V_{11} > V_{12}$) while the opposite is true for customers type 2 (i.e., $V_{21} < V_{22}$).*

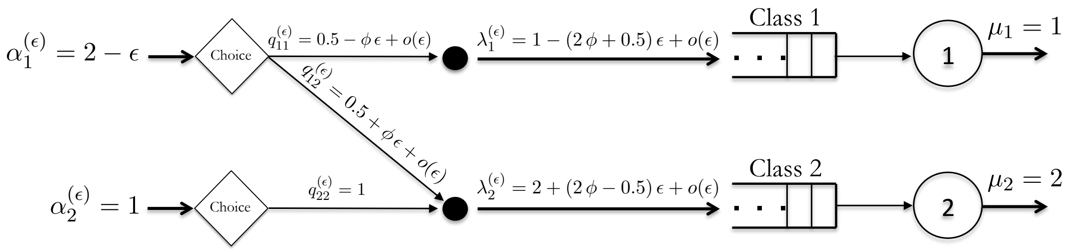
$\alpha_1^{(\epsilon)} = 2 - \epsilon$
Choice
$q_{11}^{(\epsilon)} = 0.5 - \phi\,\epsilon + o(\epsilon)$
$q_{12}^{(\epsilon)} = 0.5 + \phi\,\epsilon + o(\epsilon)$
$\lambda_1^{(\epsilon)} = 1 - (2\,\phi + 0.5)\,\epsilon + o(\epsilon)$
Class 1
1
$\mu_1 = 1$
$\alpha_2^{(\epsilon)} = 1$
Choice
$q_{22}^{(\epsilon)} = 1$
$\lambda_2^{(\epsilon)} = 2 + (2\,\phi - 0.5)\,\epsilon + o(\epsilon)$
Class 2
2
$\mu_2 = 2$

Figure 5: Example with two customer types, two service classes and two servers.

For the given values of the arrival and service rates as well as the preferences of the customers, it should be intuitively clear that an equilibrium strategy $q^{(\epsilon)}$ for the ϵ^{th} system takes the form $q_{22}^{(\epsilon)} = 1 - q_{21}^{(\epsilon)} = 1$ and $q_{12}^{(\epsilon)} = 1 - q_{11}^{(\epsilon)} = 0.5 + \phi\,\epsilon + o(\epsilon)$ for some scalar ϕ such that $|\phi| < 1/4$ (this condition ensures that the queueing system is stable for $\epsilon >$ small enough). The strategy profile $q^{(\epsilon)}$ converges, as $\epsilon \downarrow 0$, to $\hat{q}^*$ given by $\hat{q}_{11}^* = \hat{q}_{12}^* = 1/2$ and $\hat{q}_{22}^* = 1 - \hat{q}_{21}^* = 1$. Thus, in the limit, half of type 1 customers are served by server 1 and all type 2 customers are served by server 2.

Since each service class behaves as an M/M/1 queue, the scaled steady-state waiting times in the ϵ^{th} system are given by

$$\widehat{W_1}^{(\epsilon)}(q^{(\epsilon)}) = \frac{1}{0.5 + 2\,\phi + O(\epsilon)} \qquad \text{and} \qquad \widehat{W_2}^{(\epsilon)}(q^{(\epsilon)}) = \frac{1}{0.5 - 2\,\phi + O(\epsilon)}.$$

It follows from the above that to characterize the limiting value of the waiting times the limiting strategy $\hat{q}^*$ is not enough, and the direction ϕ of convergence of the strategy profile $q^{(\epsilon)}$ is necessary. To pinpoint the precise value of ϕ that will ensure that $\hat{q}^*$ is a heavy traffic equilibrium we must impose the best-response condition in Definition 7. In this example, customers type 1 randomize between service classes 1 and 2 and so they must be indifferent between them. It follows that

$$\lim_{\epsilon \downarrow 0} \big(\widehat{W_1}^{(\epsilon)}(q^{(\epsilon)}) - \widehat{W_2}^{(\epsilon)}(q^{(\epsilon)})\big) = \frac{V_{11} - V_{12}}{\delta}.$$

Letting $\beta := (V_{11} - V_{12})/\delta$, we get that choosing

$$\phi^* = \frac{2 - \sqrt{4 + \beta^2}}{4\,\beta}$$

ensures that $\hat{q}^*$ is indeed a heavy traffic equilibrium in the sense of Definition 7. □

3.5 Pareto Improvement and Chained DAGs

Consider an admissible menu in heavy traffic $M \in \widehat{\mathcal{M}}$ with a heavy traffic equilibrium $(\hat{q}^*, \widehat{W}^*, \hat{p}^*)$ and let $\mathbb{C} = \{\mathbb{C}_1, , \dots, \mathbb{C}_K\}$ be its corresponding collection of CRP components. Our next result shows that under fairly general conditions we can always find another menu with a heavy traffic equilibrium with the same collection of CRP components that (weakly) Pareto dominates $(\hat{q}^*, \widehat{W}^*, \hat{p}^*)$.

Proposition 5. *Consider an admissible menu $M \in \widehat{\mathcal{M}}$ with a heavy traffic equilibrium $(\hat{q}^*, \widehat{W}^*, \hat{p}^*)$ and CRP components $\mathbb{C} = \{\mathbb{C}_1, \dots, \mathbb{C}_K\}$. Denote by $\widehat{W}_{\mathbb{C}_k}$ the limiting scaled waiting time of component $\mathbb{C}_k$ for $k \in [K]$ and assume (after relabeling if necessary) that $\widehat{W}_{\mathbb{C}_1} \leq \widehat{W}_{\mathbb{C}_2} \leq \cdots \leq \widehat{W}_{\mathbb{C}_K}$. Suppose that $1/|a| \leq \widehat{W}_{\mathbb{C}_1} < \widehat{W}_{\mathbb{C}_2}$, then there exists a menu $M' \in \widehat{\mathcal{M}}$ with a heavy traffic equilibrium $(\hat{q}^*, \widehat{W}', \hat{p}^*)$ with the same set of CRP components $\mathbb{C}$ and such that $\widehat{W}' \leq \widehat{W}^*$. Furthermore, in this new equilibrium the CRP components in $\mathbb{C}$ are connected through a chained DAG (see Definition 6).*

Proposition 5 is significant as it reveals that for the purpose of finding an optimal service menu we can essentially restrict ourself to menus that induce a heavy traffic equilibrium with CRP components connected by a chained DAG. We will take full advantage of this property in Section 6, where we study the class of Partition service menus. We also note that we can extend the result in the proposition to include the degenerate case $1/|a| < \widehat{W}_{\mathbb{C}_1} = \widehat{W}_{\mathbb{C}_2}$[¶]. In this case, however, we can only show that for any $\varepsilon > 0$ (small) there exists a ε-heavy-traffic equilibria with the same CRP components connected by a chained DAG that (weakly) Pareto dominates $(\hat{q}^*, \widehat{W}^*, \hat{p}^*)$.

[¶] Recall that by Corollary 1 the case $1/|a| = \widehat{W}_{\mathbb{C}_1} = \widehat{W}_{\mathbb{C}_2}$ is not possible.

4 Service Menus with Two Servers

In this section we illustrate the heavy traffic analysis in the context of a system with two servers (i.e., $m = 2$). In this setting, we are able to obtain a complete solution and derive a number of insights that we will use later to analyze the general case with an arbitrary number of servers. The two-server model is also worth studying in its own right as it provides a parsimonious framework that allows for a non-trivial segmentation of service (e.g., high vs. low quality). We note that it is possible to analyze the two-server model under non heavy traffic conditions with a fair amount of detail, however, the analysis does not extend in any useful way to the general case with $m > 2$. For completeness, in Appendix B we characterize consumers' equilibrium strategy profiles for the two-server case under general traffic conditions.

Before we begin studying the possible menus and their equilibria, it helps to establish some *benchmarks* for what performance one might aim for along the dimensions of average waiting time and matching reward, respectively. Looking first at average waiting time, it is quite straightforward to see that one can not expect an average delay smaller than that of a single server queue with service rate equal to the total service rates of the m servers, and the arrival rate equal to the total arrival rate of the customer types. Under heavy-traffic, we denote this ideal scaled delay as $\overline{W}_{\text{min}}$:

$$\overline{W}_{\text{min}} = \frac{1}{|a|}. \tag{8}$$

(Recall that $|a| = \sum_{\theta\in\Theta} a_\theta$ is the aggregated capacity slack.) Next, turning to matching reward, the following max-flow linear program solves the matching that a central planner would like to implement if she had complete control over the assignment of customers to servers and were only concerned with maximizing matching rewards.

$$\overline{V}_{\text{max}} := \max_{f_{\theta j}\geq 0} \sum_{\theta,j} f_{\theta j}\, V_{\theta j} \quad \text{subject to} \quad \sum_j f_{\theta j} = A_\theta,\ \forall \theta \in \Theta \quad \text{and} \quad \sum_\theta f_{\theta j} = \mu_j,\ j = 1, 2. \tag{9}$$

It thus follows that $\overline{V}_{\text{max}}$ is an upper bound on the cumulative matching value that can be achieved by any menu in equilibrium.

We now consider the space of admissible menus. With two servers, there are three possible service classes, namely, Class 1 served only by server 1, Class 2 served only by server 2, and Class 3 served by both servers. With these three classes available, the service provider can offer one of the following five admissible service menus (see Figure 1):

- Dedicated menu (D), in which Classes 1 and 2 are offered,
- Single-line menu (SL), in which only service Class 3 is offered,
- Full menu (F), in which all three classes are offered,
- N_i menu, in which Classes i and 3 are both offered, for $i = 1, 2$.

When there are only two servers, we can index the customer types according to their relative preferences over the two servers. Specifically, we order the customer types $\{\theta_1, \theta_2, \ldots, \theta_{|\Theta|}\}$ such that $\Delta V_{\theta_i} \leq \Delta V_{\theta_j}$ for all $1 \leq i < j \leq |\Theta|$, where $\Delta V_{\theta_i} = V_{\theta 2} - V_{\theta_1}$. Let us define the subsets $\Theta_0 := \{\theta \in [\Theta] \colon \Delta V_\theta = 0\}$, $\Theta_1 := \{\theta \in [\Theta] \colon \Delta V_\theta < 0\}$, $\Theta_2 := \{\theta \in [\Theta] \colon \Delta V_\theta > 0\}$ so that customers in Θ_0 are indifferent between the two servers while customers in Θ_i strictly prefer server $i = 1, 2$. We also define $A_i := \sum_{\theta \in \Theta_i} A_\theta$ to be the limiting arrival rate of customers in Θ_i, for $i = 0, 1, 2$.

To fix ideas and notation, let us assume that the capacity of server 1 is insufficient to serve all customers who strictly prefer server 1 over server 2, i.e., $A_1 > \mu_1$. The case $A_2 > \mu_2$ is of course equivalent after relabeling the servers. The case $A_i < \mu_i$ for $i = 1, 2$ is discussed at the end of this section in Remark 4. Finally, the boundary case $A_i = \mu_i$ for $i = 1, 2$ can be analyzed using similar ideas and, for brevity, is omitted.

Under the assumption $A_1 > \mu_1$, Figure 6 depicts an example of the performance of the heavy traffic equilibrium of the five menus in the delay vs. reward quadrant. As we can see from the figure, we can split the five menus into two groups:

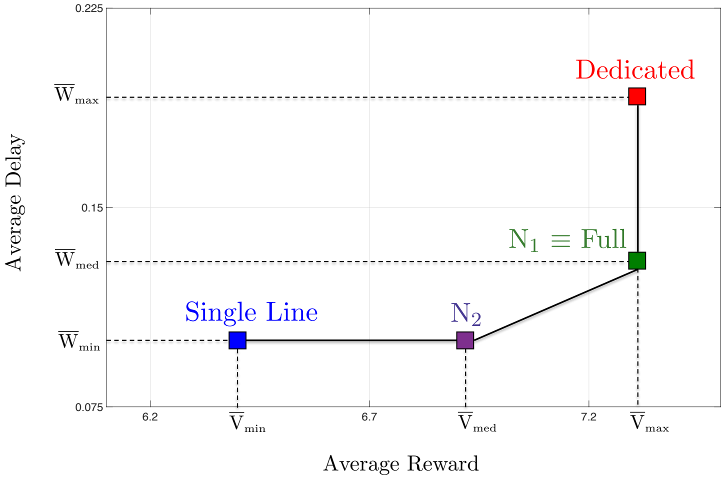

Figure 6: Performance of the heavy traffic equilibrium for Dedicated, Single Line, Full, N_1 and N_2 menus. DATA: $|\Theta| = 5$, $A = (1, 1, 3, 3, 2)$, $a = (2, 2, 2, 2, 2)$, $V_{\cdot 1} = (10, 10, 5.1, 9, 2)$, $\delta = 1$, $V_{\cdot 2} = (2, 8, 5, 10, 4)$ and $\mu_1 = 3$, $\mu_2 = 7$.

1. **Delay Minimizing Menus:** The Single Line and N_2 menus achieve the minimum possible average scaled waiting time, $\overline{W}_{\min}$.

2. **Reward Maximizing Menus:** The Dedicated, Full and N_1 menus all lead to equilibria that attain maximum possible matching reward, $\overline{V}_{\max}$. Furthermore, the equilibrium of the Full and N_1 turn out to be equivalent in heavy traffic.

To get some intuition about this segmentation of the menus, consider Figure 7 that summarizes the heavy traffic equilibrium outcome for the five menus in terms of matching flows and corresponding DAG of CRP components. Note that both the Single Line and the N_2 menu induce a single CRP component in equilibrium. For the Single Line this is trivially the case and for the N_2 menu this follows from the fact that $A_1 > \mu_1$ and so there are enough customers who want to join class 3 to ensure a positive flow from class 3 to server 2 in equilibrium. Thus, with a single CRP component, Corollary 1 implies that customers' average waiting time is minimized and equals $\overline{W}_{\min}$. In terms of matching rewards, however, the Single Line and N_2 menus have different performance. On one hand,

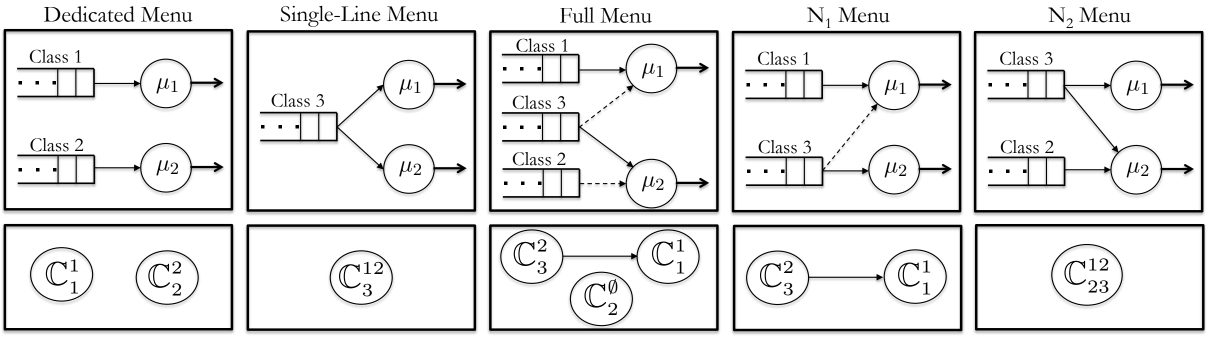

Figure 7: Summary of the heavy traffic equilibrium outcome for the five menus in terms of matching flows and DAG of CRP components.
Top Panel depicts the matching flows between the customer types and service classes indicate equilibrium strategies. Solid (dashed) arrows between the service classes and servers indicate asymptotically non-negligible (negligible) flows.
Bottom Panel depicts the DAG that emerges in the heavy traffic equilibrium, where $\mathbb{C}_{\mathcal{C}}^{\mathcal{S}}$ denotes a CRP that includes customer classes in $\mathcal{C}$ and servers in $\mathcal{S}$.

the Single Line produces the lowest average reward ($\overline{V}_{\min}$ in Figure 6) among all five menus, while the N_2 menu generates an intermediate reward value $\overline{V}_{\text{med}}$ in Figure 6. To compute the value of $\overline{V}_{\min}$ note that in the Single Line menu all customers –irrespective of their type– are matched to servers in proportion to their service rate, that is,

$$\overline{V}_{\min} = \sum_{\theta\in\Theta} \frac{A_\theta}{|A|}\left(\frac{\mu_1}{\mu_1+\mu_2} V_{\theta 1} + \frac{\mu_2}{\mu_1+\mu_2} V_{\theta 2}\right). \tag{10}$$

On the other hand, to compute $\overline{V}_{\text{med}}$, we note that since the N_2 menu induces a single CRP component, the two service classes offered in N_2 (namely, classes 2 and 3) have the same waiting time in equilibrium. As a result, all customer types that strictly prefer server 1 join class 3 and all customer types that strictly prefer server 2 join class 2. Customers who are indifferent between the two servers are also indifferent between the two service classes since they have the same waiting time. However, the assumption $A_1 > \mu_1$ together with our 'optimistic' formulation (see Remark 1) imply that all these indifferent customers join class 2 in equilibrium. It follows that

$$\overline{V}_{\text{med}} = \sum_{\theta\in\Theta_1} \frac{A_\theta}{|A|}\left(\frac{\mu_1}{A_1} V_{\theta 1} + \frac{A_1-\mu_1}{A_1} V_{\theta 2}\right) + \sum_{\theta\in\Theta_0\cup\Theta_2} \frac{A_\theta}{|A|} V_{\theta 2}. \tag{11}$$

Let us turn to the three reward maximizing menus: Dedicated, Full and N_1. The common feature of these three menus is that they all include service Class 1 and since $A_1 > \mu_1$: *(i)* server 1 exclusively serves customer types that prefer it the most, leading to reward maximization, and *(ii)* there is at least one customer type in Θ_1 that must be indifferent between joining Class 1 and some other class. It is precisely this indifference condition that pinpoints the heavy traffic equilibrium for these three menus. In terms of the average delay experienced by customers in equilibrium, the Full and N_1 menus produce the same average delay $\overline{W}_{\text{med}}$, while the Dedicated produces an average delay $\overline{W}_{\max}$, with $\overline{W}_{\text{med}} \le \overline{W}_{\max}$. This is an example of the Pareto improvement described in Proposition 5, since the DAGs of Full and N_1 menu can be seen as chaining the CRP components of the Dedicated menu.

The following proposition summarizes the performance of the heavy traffic equilibrium for each of the five menus.

Proposition 6. *Suppose that $A_1 \geq \mu_1$. Then, the performance of the heavy traffic equilibrium, in terms of average waiting times and matching rewards, for each of the five menus is given by:*

	Dedicated	N_1	Full	N_2	Single-line
Avg. Waiting Times	$\overline{W}_{\text{max}}$	$\overline{W}_{\text{med}}$	$\overline{W}_{\text{med}}$	$\overline{W}_{\text{min}}$	$\overline{W}_{\text{min}}$
Avg. Matching Rewards	$\overline{V}_{\text{max}}$	$\overline{V}_{\text{max}}$	$\overline{V}_{\text{max}}$	$\overline{V}_{\text{med}}$	$\overline{V}_{\text{min}}$

such that $\overline{W}_{min} \leq \overline{W}_{med} \leq \overline{W}_{max}$ and $\overline{V}_{min} \leq \overline{V}_{med} \leq \overline{V}_{max}$. The values of $\overline{W}_{min}$, $\overline{V}_{min}$, $\overline{V}_{med}$ and $\overline{V}_{max}$ are derived in equations (8)-(9) and the values of $\overline{W}_{med}$ and $\overline{W}_{max}$ are derived in the proof of the proposition in equations (A5) and (A4), respectively.

Let us conclude this section with the following two remarks.

Remark 4. (First Best Menu). *There are two cases in which the service provider can achieve a first outcome, namely, $\overline{W}_{min}$ delays and $\overline{V}_{max}$ rewards:*

(i) *Suppose $A_i = \mu_i$ for either $i = 1$ or $i = 2$. Then, offering the N_{3-i} menu achieves first best. To see this, take for example the case $A_1 = \mu_1$, then we get from (11) that $\overline{V}_{med} = \overline{V}_{max}$ and from Proposition 6 we conclude that the N_2 menu Pareto dominates the other four menus as it achieves the best performance in both dimensions (waiting times and rewards).*

(ii) *Consider the case $A_i < \mu_i$ for $i = 1, 2$, that is, when both servers have excess capacity to serve the customer types that strictly preferred them. In this case, the Full menu is optimal as it Pareto dominates the other four menus. To see this, note that the condition $A_i < \mu_i$ implies that a stable strategy is to have customers in Θ_i joining class i (for $i = 1, 2$) and the indifferent customers in Θ_0 joining class 3. This strategy will naturally maximize average matching rewards. Furthermore, in the heavy traffic regime, this strategy will induce a single CRP component and so the average waiting time of each service class is the same. Thus, no customer class has an incentive to switch to another class. As a result, a Full menu achieves simultaneously the minimum average waiting time and the maximum matching reward and it is therefore optimal.*

We also note that the equivalence between the Full and N_1 menus does not hold anymore when $A_i < \mu_i$ for $i = 1, 2$. In this case, the N_1 menu does not produce maximum matching rewards since some customers in Θ_2 will have to be served by server 1 in equilibrium. ⋄

Remark 5. (Trivial CRP Components) *In Figure 7, the DAG induced by the Full menu has a CRP component $\mathbb{C}_2^{\emptyset}$ which includes class 2 and no server. This anomaly happens because even though class 2 is offered there is no customers joining this class in the heavy traffic equilibrium. Note that despite the fact that there is no flow of customers joining class 2, we still need to assign a waiting time to this class to enforce equilibrium conditions. Caldentey et al., 2022 provide a detailed discussion of how to compute the waiting time of these trivial CRP components in heavy traffic.* ⋄

5 First Best Menus

We saw in our discussion of the two-server case that it is sometimes possible to offer a service menu that achieves a first best outcome, that is, maximum possible matching values and minimum possible average waiting times simultaneously (see Remark 4). In this section, we investigate conditions under which a first best menu exists in a system with an arbitrary number of servers. To this end, we find it convenient to first discuss two special menus, namely the Single Line (SL) and Dedicated (D) menus, which exhibit extreme and contrasting performance in terms of matching rewards and waiting times. While a Single Line menu minimizes waiting times at the expense of matching values the opposite is true for the Dedicated menu. This is illustrated in Figure 6 for the two-server case.

5.1 Single Line Menu

In the Single Line menu the service provider offers a single service class which is served by all servers. This is the simplest and most common service configuration used in practice in which all m servers serve a single service class. By Corollary 1, the Single Line exhibits complete resource pooling and therefore minimizes average waiting times in heavy traffic, $\widehat{W}^{\mathrm{SL}} = 1/|a|$. Thus, it is an optimal menu when the service provider is interested in minimizing customers' average waiting times exclusively (i.e., $\zeta = \infty$). However, as we saw in the two-server model, the Single-Line menu is not Pareto optimal in general. Actually, our next result reveals that while the Single-Line menu minimizes waiting times, it also minimizes average matching rewards.

Theorem 3. *For any an admissible menu in heavy traffic $M \in \widehat{\mathcal{M}}$ and any heavy traffic equilibrium strategy profile $\hat{q}^* \in \widehat{\mathcal{Q}}^*(M)$ under M, let $\overline{V}(M, \hat{q}^*)$ be the average matching rewards under the pair $(M, \hat{q}^*)$. Let also $\overline{V}^{\mathrm{SL}}$ be the average matching reward under the Single Line menu. Then,*

$$\overline{V}^{\mathrm{SL}} \leq \overline{V}(M, \hat{q}^*).$$

The key limitation of the Single Line menu is its inability to customize the matching between customers and servers since all customers are essentially treated equally. This raises the question of how to design a menu that maximizes customer's rewards among all menus that have an equilibrium with a single CRP component. We will return to this question in Section 7.2.

5.2 Dedicated Menu

In the Dedicated menu each server operates independently serving its own service class. In other words, the matching topology $M^{\mathrm{D}} = [m_{ij}^{\mathrm{D}}] \in \{0,1\}^{m \times m}$ of the Dedicated menu satisfies $m_{ij}^{\mathrm{D}} = \mathbb{1}(i = j)$. In contrast to the Single Line menu, the Dedicated menu has no resource pooling but offers full flexibility to match customers to servers, and in Theorem 4 below we show that this matching flexibility is actually maximal. To this end, let us consider the following max-flow problem for system ϵ, which is central to our characterization of first best menus:

$$\begin{aligned} \overline{\mathcal{V}}^{(\epsilon)} := \max_{f_{\theta j}^{(\epsilon)} \geq 0} & \sum_{\theta, j} f_{\theta j}^{(\epsilon)} V_{\theta j} && && \textbf{(Max-flow)} \\ \text{subject to} \quad & \sum_{j} f_{\theta j}^{(\epsilon)} = \alpha_{\theta}^{(\epsilon)} && \forall \theta \in [\Theta], && \textsf{(flow balance)} \\ & \sum_{\theta} f_{\theta j}^{(\epsilon)} \leq \mu_j && \forall j \in [m], && \textsf{(capacity)} \end{aligned}$$

where $f_{\theta j}^{(\epsilon)}$ represents the flow of customers type θ served by server j. We note that the value of $\overline{\mathcal{V}}^{(\epsilon)}$ corresponds to the maximum average matching reward that a central planner can achieve if she has full control on how to match customers to servers. It follows that $\overline{\mathcal{V}}^{(\epsilon)}$ provides an upper bound on the maximum matching reward that the service provide can get from any equilibrium. Interestingly, our next result shows the Dedicated menu achieves this upper bound asymptotically. In other words, the Dedicate menu is an optimal menu if the service provider is completely insensitive to waiting times (i.e., $\zeta = 0$). This is interesting because customers' equilibrium strategies still depend on the waiting times of each service class, and so even if $\zeta = 0$ the service provider cannot simply disregard the effect of waiting times on the overall performance.

Theorem 4. *Let $V^{(\epsilon)}$ be the matching value of an equilibrium for the Dedicated menu for system ϵ. Then, $\overline{\mathcal{V}}^{(\epsilon)} - V^{(\epsilon)} = \mathcal{O}(\epsilon)$, i.e., the* Dedicated *menu asymptotically maximizes average matching value in heavy traffic.*

Proof Sketch: Since some elements of the proof are quite insightful and useful for the discussion that follows, we provide a quick proof sketch here and defer a full version to the Appendix. First, let us introduce the dual variables $\eta_\theta^{(\epsilon)}$ for the flow balance for customer type θ, and $\omega_j^{(\epsilon)}$ for the capacity constraint for server j. The dual problem to (**Max-flow**) is

$$\min_{\omega_j^{(\epsilon)} \geq 0, \eta_\theta^{(\epsilon)}} \sum_\theta \alpha_\theta^{(\epsilon)} \eta_\theta^{(\epsilon)} + \sum_j \mu_j \, \omega_j^{(\epsilon)} \quad \text{subject to} \quad \eta_\theta^{(\epsilon)} + \omega_j^{(\epsilon)} \geq V_{\theta j} \qquad \forall \theta, j. \qquad \textbf{(Dual-Max-flow)}$$

The main idea is to show that any equilibrium strategy $q^{(\epsilon)} \in \mathcal{Q}^*(M^{\mathrm{D}})$ for the Dedicated menu (which exists due to Theorem 1) induces a vector of flow rates from customers to servers, $f_{\theta j}^{(\epsilon)}(q^{(\epsilon)})$, that can be used to construct a feasible dual solution such that approximate complementary slackness holds in the following sense:

$$\left(\mu_j - \sum_\theta f_{\theta j}^{(\epsilon)}\right) \omega_j^{(\epsilon)} = \mathcal{O}(\epsilon), \qquad \text{and} \qquad \left(\eta_\theta^{(\epsilon)} + \omega_j^{(\epsilon)} - V_{\theta j}\right) f_{\theta j}^{(\epsilon)} = 0,$$

for all θ and j, which then guarantees that $f_{\theta j}^{(\epsilon)}$ is approximately optimal for (**Max-flow**) with an $\mathcal{O}(\epsilon)$ additive suboptimality, which vanishes as $\epsilon \downarrow 0$.

In particular, given that the expected waiting time at queue j under $f_{\theta j}^{(\epsilon)}$ equals $\widehat{W}_j^{(\epsilon)} = 1/(\mu_j - \sum_\theta f_{\theta j}^{(\epsilon)})$, we define for all $j \in [m]$ and $\theta \in \Theta$,

$$\omega_j{}^{(\epsilon)} = \delta \, \widehat{W}_j^{(\epsilon)} \qquad \text{and} \qquad \eta_\theta{}^{(\epsilon)} = \max_j \left\{ V_{\theta j} - \omega_j{}^{(\epsilon)} \right\} \tag{12}$$

as a feasible dual solution. By the definition above, $\eta_\theta^{(\epsilon)}$ is in fact the utility of type θ customers. To see the intuition behind why complementary slackness holds, the first set of conditions follow from the definition of $\omega_j^{(\epsilon)}$. For the second set, since under any equilibrium, $f_{\theta j}^{(\epsilon)} > 0$ only if $V_{\theta j} - \omega_j^{(\epsilon)} \geq \eta_\theta^{(\epsilon)}$, exact complementary slackness holds for the second set of conditions. □

Remark 6. *As we alluded to in the proof sketch, the optimal dual variables of* (**Max-flow**) *(for the limiting case $\epsilon = 0$) have the following interpretation: ω_j denotes the limiting scaled mean delay disutility for server j under the Dedicated menu, and η_θ denotes the average utility of customer type θ under delay disutilities $\{\omega_j\}$. However, the dual solution is only determined up to a translation; for any τ, $(\eta_\theta + \tau)$ and $(\omega_j - \tau)$ are also an optimal dual solution. Therefore without loss of generality, we can assume $\min_j \omega_j = 0$, so that the true scaled delay disutility for server j is $\delta \widehat{W}_j = \omega_j + \omega_0$ for some ω_0.*

Our next somewhat surprising result generalizes Theorem 4 in the sense that any menu M which includes the Dedicated menu as a sub-menu also maximizes the average matching value.

Theorem 5. *Let M be any service menu which includes the Dedicated menu as a submenu. That is, for every server j, there exists a service class $i(j)$ such that $m_{i(j),j} = 1$, and $m_{i(j),j'} = 0$ for any $j' \neq j$. Then the menu M attains the maximum matching reward under any heavy-traffic equilibrium.*

5.3 Necessary and Sufficient Conditions for First Best Outcomes

From the performance of the Single Line and Dedicated menus we have that for a menu to achieve first best it must simultaneously induce an equilibrium with (i) a single CRP component and (ii) matching flows that coincide with the solution of the (**Max-flow**) problem. In this section, we use this insight to identify necessary and sufficient conditions for a first best outcome to be achievable.

Theorem 6. (Necessary Conditions) *If the service provider is able to achieve a first best outcome, then there exists a solution to* (**Max-flow**) *with $\epsilon = 0$ such that the following two conditions hold:*

1. *The arcs associated with strictly positive flows form a connected graph.*
2. *Every customer type weakly prefers their matching outcome to that of any other customer type.*

A proof of this theorem can be found in an appendix, we will briefly provide some intuition here. If a first best outcome can be achieved, then the flows between service classes and servers must form a connected graph to support a single CRP component. This implies that the flows between customer types and servers must also form a connected graph. Similarly, since a first best outcome necessarily achieves the maximum possible matching values, we know that the flows between customer types and servers (via the service classes offered) must also be a solution to (**Max-flow**). Since the flows form an equilibrium, we know that no customer type prefers the matching outcome of any other customer type.

One circumstance in which it is not possible to satisfy these conditions is if the solution to (**Max-flow**) is such that every server j has some customer type θ that is only being served by server j, and there are no indifferent customers. This might occur if there is an agreed upon ranking over servers between customer types, and the arrival rate of each customer type is less than the service capacity of each server. In this case, the only way to achieve a maximum reward outcome is to offer the dedicated menu as a sub-menu. However, if the dedicated menu is being offered as a sub-menu, if service times are equal across service classes, all customer types will want to join the service class being served by their most preferred server. So there is no equilibrium in which the (**Max-flow**) rewards and minimum average delays are achieved simultaneously.

Next, we provide sufficient conditions for a first best outcome to be achievable.

Theorem 7. (Sufficient Conditions) *The service provider is always able to achieve a first best outcome if there exists a solution to* (**Max-flow**) *with $\epsilon = 0$ such that the following to conditions hold:*

1. *The basic feasible activities induce a connected tree*
2. *Every customer type weakly prefers their matching outcome to that of any other customer type.*

The menu that achieves the first best outcome is the menu in which there is a single service class for each customer type, and that service class consists of all of the servers that they are connected to in the Max Flow solution.

As the proof of this theorem is straightforward and intuitive, we include it here.

Proof. To see that these conditions are sufficient, we can consider what would happen if the proposed menu were offered. Since the flows form a connected tree, we know that these flows are those that would be achieved if this menu were to be offered, and each customer class were to join their assigned service class. As the resulting graph is connected, we know that a single CRP component is achieved, and hence minimum possible expected waiting times occur. Because each customer type weakly prefers their own matching outcomes to that of any other customer type, and waiting times across service classes are equal, it is an equilibrium for each customer type to join their assigned service class. □

This also provides some intuition as to why the conditions stated in Theorem 6 are not sufficient. Should the flows associated with the feasible activities form a graph with cycles, we cannot guarantee that the flows can be achieved by any particular menu as we can should the flows form a tree.

One way that these sufficient conditions can be satisfied is to have 'enough' indifferent customers in the system. By this we mean that there is some mass of customer types who have strict preferences between servers, and some mass of customers who are indifferent between pairs of servers. If there is enough service capacity so that all customer types with strict preferences can be served by their most preferred servers, and an ordering of servers so that for every pair of servers $(j, j+1)$ for $j = 1, ..., m-1$ there is a customer type who is indifferent between servers j and $j+1$, then the sufficient conditions will be satisfied.

6 Partition Menus

In the previous section we identified conditions under which there exists a menu that achieves first best outcome. In general, however, first best cannot be achieved and an optimal menu must appropriately balance the trade-off between waiting times and matching rewards. In this section, we investigate this trade-off by restricting ourselves to the study of a special class of *Partition menus* in which the set of servers is partitioned into K pools $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_K\}$ for some $K \in [m]$. We will consider two classes of partition menus.

- Pure Partition Menus: These are menus in which each partition of servers $\mathcal{S}_k$ is dedicated to serving exclusively a single service class, say $\mathcal{C}_k$. The left panel in Figure 8 depicts an example of a pure partition menu with four customers classes and five servers. In this example, servers are partitioned into two sets $\mathcal{S}_1 = \{1, 2, 3\}$ and $\mathcal{S}_2 = \{4, 5\}$ with all servers in partition $\mathcal{S}_i$ serving exclusively service class i, for $i = 1, 2$.

  It is worth noticing that in heavy traffic, a partition menu consists of K disconnected CRP components $\mathbb{C} = \{\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_K\}$, with $\mathbb{C}_k = (\mathcal{C}_k, \mathcal{S}_k)$.

- Chained Partition Menus: These are modified pure partition menus with some additional connectivity among the CRP components $\mathbb{C} = \{\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_K\}$ so that the underlying DAG has a chained structure (see Definition 6). Thus, every chained partition menu has associated an underlying pure partition menu that defines it. For example, the right panel in Figure 8 depicts a chained partition menu associated to the pure partition in the right panel that includes a link (dashed arc) connecting $\mathbb{C}_1$ to $\mathbb{C}_2$.

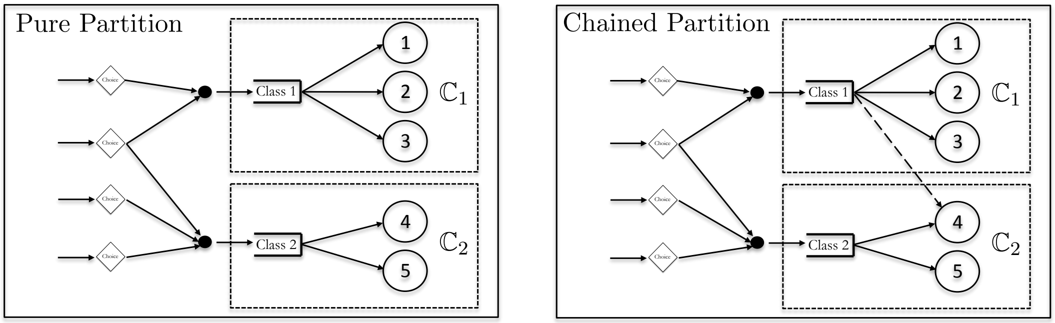

Figure 8: Example of pure partition and chained partition menus with two partitions of servers $\mathcal{S}_1 = \{1,2,3\}$ and $\mathcal{S}_2 = \{4,5\}$.

While restrictive, partition menus have a number of desirable properties from a practical standpoint as they are easy to explain to customers and require limited scheduling coordination among the servers. For instance, in a pure partition menu each server can manage FCFS requirements by tracking a single service class and customers only need to know their queueing position in a single line to assess their service status. In addition, by varying the numbers of partitions and their composition, partition menus offer a fair amount of flexibility that the service provider can use to trade-off matching rewards and waiting times. For instance, two notable examples of partition menus are the Single Line and the Dedicated menu discussed in Sections 5.1 and 5.2, respectively.

6.1 Pure Partition Menus

Let us fix a partition $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_K\}$, with all the servers in partition $\mathcal{S}_k$ serving a unique service class $\mathcal{C}_k$. It is easy to see that each pair $\mathbb{C}_k = (\mathcal{C}_k, \mathcal{S}_k)$ corresponds to a different CRP component in any heavy traffic equilibrium. In this setting, a strategy profile can be represented by a matrix $q = [q_{\theta k}]$, where $q_{\theta k}$ is the probability that a type θ customer joins $\mathcal{C}_k$. Moreover, since each service class $\mathcal{C}_k$ is served exclusively by the servers in $\mathcal{S}_k$, the limiting matching probabilities $\hat{p}$ of any heavy traffic equilibrium must trivially satisfy $\hat{p}_{kj} = \mathbb{1}(j \in \mathcal{S}_k)\, \mu_j / \mu_{\mathcal{S}_k}$, where $\mu_{\mathcal{S}_k} := \sum_{j \in \mathcal{S}_k} \mu_j$. It follows that the average limiting reward that a type θ customer gets from joining service class $\mathcal{C}_k$ equals

$$\overline{V}_{\theta k} := \sum_{j \in \mathcal{S}_k} \frac{\mu_j \, V_{\theta j}}{\mu_{\mathcal{S}_k}}.$$

It is not hard to see that a pure partition menu with servers' partition $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_K\}$ behaves, in the heavy traffic limit, as the Dedicated menu in which each partition of servers $\mathcal{S}_k$ acts as a '*super-server*' with capacity $\mu_{\mathcal{S}_k}$ and with a matrix of matching rewards $\overline{V} = [\overline{V}_{\theta k}]$ between customers types and super-servers. With this interpretation, on can show that Theorem 4 extends to this case in a relatively straightforward fashion. Specifically, consider the following modified version of the max-flow problem parameterized by the partition $\mathcal{S}$:

$$\overline{\mathcal{V}}_{\mathcal{S}}^{(\epsilon)} := \max_{f_{\theta k}^{(\epsilon)} \geq 0} \sum_{\theta,k} f_{\theta k}^{(\epsilon)} \, \overline{V}_{\theta k} \quad \text{subject to} \quad \sum_{k} f_{\theta k}^{(\epsilon)} = \alpha_{\theta}^{(\epsilon)} \quad \text{and} \quad \sum_{\theta} f_{\theta k}^{(\epsilon)} \leq \mu_{\mathcal{S}_k}, \tag{13}$$

where $f_{\theta k}^{(\epsilon)}$ represents the flow of customers type θ joining service class $\mathcal{C}_k$. As in the case of the Dedicated menu, $\overline{\mathcal{V}}_{\mathcal{S}}^{(\epsilon)}$ provides an upper bound on the maximum matching reward that the service provide can get from any equilibrium under the pure partition menu $\mathcal{S}$.

Corollary 2. *The average matching value $V_{\mathcal{S}}^{(\epsilon)}$ of any equilibrium for the pure partition menu with server partition $\mathcal{S}$ satisfies $\overline{\mathcal{V}}_{\mathcal{S}}^{(\epsilon)} - V_{\mathcal{S}}^{(\epsilon)} = \mathcal{O}(\epsilon)$.*

According to the previous result, all equilibria associated to the pure partition menu with partition $\mathcal{S}$ generate the same matching value $\mathcal{V}_{\mathcal{S}} := \min_{\epsilon \downarrow 0} \overline{\mathcal{V}}_{\mathcal{S}}^{(\epsilon)}$, in the heavy traffic limit. Under the following assumption on the max-flow problem (13), the limiting scaled waiting time of the pure partition menu is also uniquely determined.

Assumption 1. *The solution to (13) with $\epsilon = 0$ is unique, and the basic feasible activities (that is, the edges (θ, k) with $f_{\theta k} > 0$) induce a connected tree.*

Assumption 1 is quite mild. For example if one were to generate a random instance of the service system by sampling the valuations $V_{\theta j}$ from non-atomic distributions then the maximum flow is unique with probability 1. Similarly, if either the arrival rates A_{θ} or the service rates μ_j are randomly sampled from non-atomic distributions then the maximum flow forest is a connected tree with probability 1.

Under Assumption 1, the limiting mean scaled waiting times of all service classes in the pure partition menu are determined up to an additive constant. This is because a customer type θ randomizing between service classes $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ must be indifferent between them, and hence it must be true that $\overline{V}_{\theta k} - \delta \widehat{W}_k = \overline{V}_{\theta k'} - \delta \widehat{W}_{k'}$. The connectivity assumption then implies that knowing $\widehat{W}_k$ for some service class yields the waiting time for all service classes. Recall from Remark 6 that we can express the limiting scaled waiting times $\widehat{W}_k$ in terms of the dual variables ω_k for the service capacity constraints in (13). Specifically, there exist a vector of dual variables $\{\omega_k\}$ with $\min_k \omega_k = 0$ and a scalar ω_0 such that $\delta \, \widehat{W}_k = \omega_k + \omega_0$. We use this representation in the next proposition to derive the precise waiting times under a partition menu.

Proposition 7. *Suppose Assumption 1 holds and let $\{\omega_k\}$ be a vector of dual variables for the service capacity constraints in (13) such that $\min_k \omega_k = 0$. Then, the limiting scaled mean waiting times for the pure partition menu are given by $\widehat{W}_k^{\text{PB}} = (\omega_k + \omega_0)/\delta$ where $\omega_0 \geq \delta/|a|$ solves:*

$$\sum_{k=1}^{K} \frac{\delta}{\omega_k + \omega_0} = |a|.$$

We omit a formal proof as the intuition is simple: Under the pure partition menu, each service class $\mathcal{C}_k$ behaves asymptotically in the heavy traffic limit as an independent $M/M/1$ queue with service capacity $\mu_{\mathcal{S}_k}$. Thus a limiting scaled mean waiting time of $\widehat{W}_k$ implies $\lim_{\epsilon \downarrow 0} (\mu_{\mathcal{S}_k} - \lambda_k^{(\epsilon)})/\epsilon = 1/\widehat{W}_k$, where $\lambda_k^{(\epsilon)} = \sum_{\theta} f_{\theta k}^{(\epsilon)}$ is total arrival rate at service class $\mathcal{C}_k$. Further, $\lim_{\epsilon \downarrow 0} \sum_j (\mu_{\mathcal{S}_k} - \lambda_k^{(\epsilon)})/\epsilon = |a|$ by the heavy-traffic scaling in (3), which provides the necessary condition to pin down ω_0.

6.2 Chained Partition Menus

Corollary 2 shows that pure partition menus maximize matching values for a given partition of servers. At the same time, they do not allow any form of capacity sharing between partitions, and this can lead to poor performance in terms of waiting times. To partially correct for this deficiency, we exploit the result Proposition 5 and consider a modified class of pure partition menus by "chaining" the service classes in increasing order of their waiting time. We refer to this class of menus as *chained* partitions. Intuitively, while a pure partition menu results in a DAG with K disconnected CRP components (where each partition of servers along with their service class are in a CRP component of their own), a chained partition leads to a DAG which is a directed path (i.e., a single topological order), and allows for capacity pooling across CRP components. A special case of this construction is the N_1 menu in Section 4, where the chaining is apparent in Figure 7 (see also the right panel in Figure 8).

Recall that Proposition 2 provides a characterization of a class of scaled limiting waiting times that can be implemented in heavy traffic using a chained DAG. We take advantage of this result to derive the waiting times of a chained partition menu under the following additional assumption.

Assumption 2. *There exists an optimal vector $\{\omega_k\}$ of dual variables for the service capacity constraints in (13) such that $0 = \widehat{\omega}_{(1)} < \widehat{\omega}_{(2)}$.*

In the statement of the following proposition, we let $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ be a fixed partition of servers and $\mathcal{C}_k$ the service class connected to all servers in $\mathcal{S}_k$. We let $M_{\mathcal{S}}^{\mathrm{PB}}$ denote the pure partition menu defined by $\{(\mathcal{C}_k, \mathcal{S}_k) \colon k \in [K]\}$.

Proposition 8. *Let Assumptions 1 and 2 hold and let $\{\omega_k\}$ be the optimal vector of dual variables satisfying the conditions in Assumption 2. Without loss of generality, let us relabel the K service partitions in such a way that $0 = \omega_1 < \omega_2 \le \omega_3 \le \cdots \le \omega_K$. Define a chained partition menu $M_{\mathcal{S}}^{\mathrm{CB}}$ by extending the pure partition menu $M_{\mathcal{S}}^{\mathrm{PB}}$ as follows: add a link connecting service class $\mathcal{C}_k$ to any server in partition $\mathcal{S}_{k+1}$ for $k = 1, \ldots, K-1$. The resulting chained partition menu generates maximum matching value $\bar{\mathcal{V}}$ and has limiting scaled waiting times given by*

$$\widehat{W}_1^{\mathrm{CB}} = \frac{1}{|a|} \qquad \textit{and} \qquad \widehat{W}_k^{\mathrm{CB}} = \frac{\omega_k}{\delta} + \frac{1}{|a|}, \quad k = 2, \ldots, K.$$

It follows from Proposition 7 that $\widehat{W}_k^{\mathrm{CB}} = \widehat{W}_k^{\mathrm{PB}} - \widehat{W}_1^{\mathrm{PB}} + \frac{1}{|a|} \le \widehat{W}_k^{\mathrm{PB}}$. Thus, from the prespective of the service provider, the chained menu $M_{\mathcal{S}}^{\mathrm{CB}}$ (weakly) Pareto dominates the pure partition menu $M_{\mathcal{S}}^{\mathrm{PB}}$.

We note that under the chained partition menu $\widehat{W}_1^{\mathrm{CB}} = 1/|a|$, which by Corollary 1 is the delay under a completely pooled system and the lowest delay possible for a service class under any menu.

6.3 Optimal Partitions

We conclude this section by developing a mixed-integer linear program (MILP) to find an optimal chained partition menu. As these menus are constructed from partitions of servers, the number of possible menus grows rapidly with the number of servers. However, many of these menus will be Pareto dominated by others. The MILP formulation in Figure 9 assumes a fixed number K of partitions and finds the optimal partition of servers $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$. By varying the value of K from 1 to m we can find the optimal chained partition menu. We will also describe a process that uses the MILP to identify the set of Pareto efficient chained menus.

The following are the main decision variables used in the MILP formulation:

-) m_{kj}: 1 if server j belongs to partition $\mathcal{S}_k$ and 0 otherwise.
-) $f_{\theta kj}$: flow of type-θ customers joining $\mathcal{C}_k$ and served by server j.
-) $f_{\theta k}$: flow of type-θ customers joining $\mathcal{C}_k$.
-) $V_{\theta kj}$: value that a type-θ customer gets from joining class $\mathcal{C}_k$ and service from server j.
-) $V_{\theta k}$: average value that a type-θ customer gets from joining class $\mathcal{C}_k$.
-) ω_{kj}: waiting time experienced by a customer who joins $\mathcal{C}_k$ and get served by server j.
-) ω_k: waiting time experienced by a customer who joins $\mathcal{C}_k$.

Objective:

$$\mathcal{V}_K^{\mathrm{PB}} := \max \sum_{\theta kj} f_{\theta kj} \cdot V_{\theta j} - \zeta \sum_{kj} \mu_j \omega_{kj} \tag{14}$$

Constraints:

Server assignment:
$$\sum_k m_{kj} = 1, \qquad \sum_j m_{kj} \geq 1. \tag{15}$$

Enforcing max matching value:
$$\eta_\theta + \delta\, \omega_k \geq \overline{V}_{\theta k}, \quad \sum_{\theta kj} f_{\theta kj} \cdot V_{\theta j} = \sum_\theta A_\theta \eta_\theta + \delta \sum_{kj} \mu_j \omega_{kj}. \tag{16}$$

Waiting time within partitions:
$$\omega_k + (m_{kj} - 1)M \leq \omega_{kj} \leq \omega_k, \qquad \omega_{kj} \leq m_{kj} M. \tag{17}$$

Customers' valuation for partitions:
$$\overline{V}_{\theta k} + (m_{kj} - 1)M \leq V_{\theta kj} \leq \overline{V}_{\theta k}, \quad V_{\theta kj} \leq m_{kj} M, \tag{18a}$$
$$\sum_j \mu_j V_{\theta kj} = \sum_j m_{kj} \mu_j V_{\theta j}. \tag{18b}$$

Flow balance:
$$\sum_k f_{\theta k} = \alpha_\theta, \qquad \sum_{\theta k} f_{\theta kj} \leq \mu_j. \tag{19}$$

Auxiliary flow constraints:
$$f_{\theta kj} + (m_{kl} - 1)M \leq g_{\theta kjl} \leq f_{\theta kj}, \qquad g_{\theta kjl} \leq m_{kl} M, \tag{20a}$$
$$f_{\theta k} + (m_{kj} - 1)M \leq g_{\theta kj} \leq f_{\theta k}, \qquad g_{\theta kj} \leq m_{kj} M, \tag{20b}$$
$$\sum_l g_{\theta kjl} \cdot \mu_l = g_{\theta kj} \cdot \mu_j. \tag{20c}$$

Non-negativity of decision variables

$$\{f_{\theta k}\}, \{f_{\theta kj}\}, \{g_{\theta kj}\}, \{g_{\theta kjl}\}, \{V_{\theta k}\}, \{V_{\theta kj}\}, \{\omega_k\}, \{\omega_{kj}\} \geq 0 \quad \text{and} \quad \{m_{kj}\} \in \{0, 1\}. \tag{21}$$

Figure 9: MILP for finding the optimal partition menu with K partitions.

The key idea in this MILP is that since a chained partition menu acts like a chained-dedicated menu on super-servers, we can use simultaneously the primal and dual constraints corresponding to max-flow problem in (13) to ensure that customer arrival rates are consistent with an equilibrium strategy profile. As mentioned in Remark 6, the dual variables ω_k can be interpreted as waiting times for the service classes. This means that by incorporating the dual constraints and dual variables into the MILP, we are able to include both the matching values and the waiting times of the service classes into the objective function. This is captured in the set of constraints (16).

For a given value of ζ, we can solve the MILP to find the optimal maximizing chained partition menu. In addition, we can use the MILP to generate a Pareto frontier within the class of partition menus using standard multi-objective optimisation techniques for two objectives. We refer interested readers to Marler and Arora (2004) for a review of such methods.

7 Tailored Menus

In the previous section, we approached the problem of menu design by optimizing over the class of partitioned service menus in which each service classes has associated a unique and disjoint set of servers. In this section, we take an alternative perspective and use a *mechanism design* approach to tackle the problem of finding efficient menus. Specifically, we consider the class of *Tailored* menus for which $n = |\Theta|$ and every customer type is assigned[†] a single service class in equilibrium. A key advantage of tailored menus over bundle menus is that they provide more flexibility to customize the matching between customer types and servers. On the flip side, tailored menus are more complex to design and possibly less practical from an implementation standpoint.

For brevity of exposition, we will focus on two special type of tailored menus: (i) those that maximize value matching rewards and (ii) those that minimize waiting times.

7.1 Value Maximizing Tailored Menus

The results in Sections 5.2 and 6.2 establish that a chained Dedicated menu maximizes the service provider's matching value. However, this is menu has limited capacity pooling and therefore offers no guarantee of providing a good performance in terms of waiting times. To address this limitation we will formulate a MILP that minimizes waiting times over the class of tailored menus that produces maximum matching value.

Formally, we begin by solving (**Max-flow**) (under $\epsilon = 0$) to obtain the flow $[f_{\theta j}]$, which we will assume is unique and induces a connected tree by Assumption 1. Let $S_\theta := \{j : f_{\theta j} > 0\}$ denote the set of servers with non-zero flow from customer type θ in the maximum value flow. The menu design task then is to partition S_θ for each customer type θ into *service bundles* specifically intended for θ.

We will use the following notation:

- $B_\theta = 2^{S_\theta} \setminus \emptyset$ denotes all the non-empty subsets of S_θ.
- For type θ, we call a set $b \in B_\theta$ a service bundle intended for type θ, and also use it to denote the vector $b = (b_1, \ldots, b_m)$ where $b_j = 1$ if $j \in B$ and $b_j = 0$ otherwise. Note that although we associate bundle b with a subset of servers, each such bundle is also implicitly associated with a customer type. Thus we can have one subset of servers S offered as two bundles, one for customer type θ and for customer type θ'.
- $A_{\theta b} = \sum_j f_{\theta j} b_j$ denotes the total arrival rate into bundle b (if offered) from customer type θ.

[†] The notion of assigning a service class to each customer type should be understood in a *implementation theory* sense. The idea is that by appropriately designing the service menu, the service provider can guarantee that self-interested customers will end-up joining the service class that they are supposed to join.

- For $b \in B_\theta$, for any θ':

$$\overline{V}_{\theta' b} = \frac{\sum_j f_{\theta j} b_j V_{\theta' j}}{A_{\theta b}}$$

denotes the average value type θ' obtains from type θ's bundle b. Note that here we are assuming that the flow from bundle b to the servers is consistent with the maximum flow f.

Note that given the maximum flow $\{f_{\theta j}\}$, the above are constants which we will use in our MILP formulation. We explain the decision variables and the constraints of the MILP briefly next:

Decision Variable:

-) y_b, $b \in \cup_\theta B_\theta$: These binary decision variables correspond to the possible service bundles for all customer types. A value of 1 indicates the bundle is offered and 0 indicates it is not offered.

-) W_b, $b \in \cup_\theta B_\theta$: These continuous non-negative decision variables correspond to the delay of bundles measure in units of (dis)utility up to a translation so that $\min_b W_b = 0$.

-) U_θ, $\theta \in \Theta$: These continuous non-negative decision variables correspond to the utility of a type θ customer (up to a translation).

-) W'_j, $j \in [m]$: These decision variables correspond to the delay of server j (and thus of all offered bundles containing server j), measured in units of (dis)utility and determined up to a translation.

OBJECTIVE:

$$\mathcal{W}^* := \min \sum_{\theta} \sum_{b \in B_\theta} A_{\theta,b} \cdot W_{\theta,b}$$

CONSTRAINTS:

Feasibility of menu:
$$\sum_{b \in B_\theta} b_j y_b = 1. \tag{22}$$

Consistency of waiting times:
$$W'_j - (1-y_b)M \le W_b \le W'_j + (1-y_b)M, \quad W_b \le y_b M. \tag{23}$$

Utility for each type:
$$\overline{V}_{\theta,b} - W_b - (1-y_b)M \le U_\theta \le \overline{V}_{\theta,b} - W_b + (1-y_b)M. \tag{24}$$

Incentive compatibility:
$$U_{\theta'} \ge \overline{V}_{\theta',b} - W_b - (1-y_b)M. \tag{25}$$

Non-negativity of decision variables: $\{W_b\}, \{W'_j\}, \geq 0$ and $\{y_b\} \in \{0,1\}$.

Figure 10: MILP for finding tailored menu with minimum average delay under maximum total value constraint.

Constraint (22) ensures that for each customer type θ, a server j with $f_{\theta j} > 0$ in the max value flow solution is offered in exactly one bundle intended for θ. Constraint (23) ensures that (i) the delay disutility W_b of any bundle b that is offered and contains server j equals the delay disutility W'_j for server j and (ii) the delay disutility for any bundle b that is not offered is forced to 0 (so it contributes 0 to the objective). Constraint (24) ensures that the utility of a customer type θ equals the utility of any offered bundle b intended for θ. Finally, (25) ensures that the utility of type θ' is at least the utility she derives from all offered bundles (which may or may not be intended for θ').

The objective minimizes the total delay disutility. Let $\{b_1, \ldots, b_s\}$ be the bundles selected by the MILP, so that without loss of generality we can assume that $0 = W_{b_1} \leq W_{b_2} \leq \cdots \leq W_{b_s}$. Making a similar assumption as used for Proposition 8, we can show the existence of a menu and an equilibrium where the limiting scaled mean delay of bundle b is $\widehat{W}_b = \frac{W_b}{\delta} + \omega_0^*$. In other words, the objective of the MILP measures precisely the *additional delay disutility* compared to the minimum delay disutility experienced under a single CRP matching system.

Note that for every customer type θ, we need to enumerate all 2^{S_θ} bundles. Computationally this is not prohibitive if in the maximum value tree, the degree of each customer type is small. In Section 8 we present results from numerical experiments based on the MILP in Figure 10.

7.2 Delay Minimizing Tailored Menus

According to Corollary 1, any menu that induces a heavy traffic equilibrium with a single CRP component —such as the Single Line menu— minimizes customers' limiting waiting times. In this section, we discuss how to find a menu that maximizes matching values over the class of tailored menus that support a heavy traffic equilibrium with a single CRP component.

Just like in the previous section, we will formulate this problem as a mixed-integer linear program. Consider a menu M with $|\Theta| = n$ and let $\overline{V}_{\theta i}$ denote the average reward that a customer type θ is expected to receive by joining service class i. Since, $|\Theta| = n$, in what follows we will abuse notation and refer to service class $i \in [n]$ as the one targeted to customers of type $i \in \Theta$. Similarly, we will denote by A_i the limiting arrival rate at class $i \in [n]$.

To ensure the incentive compatibility of the proposed menu, the service provider would like to design the menu in such a way that (i) it induces a single CRP component and (ii) the following IC condition is satisfied.

$$\overline{V}_{ii} \geq \overline{V}_{ik}, \quad \text{for all } k \in [n]. \tag{IC}$$

To formulate the service provider problem, we will rely on the quadratic program (**QP**) to approximate the steady-state matching probabilities for a given matching topology. Specifically, we propose the mixed integer linear program (MILP)[‡] presented in Figure 11 to identify the matching probabilities that maximizes the approximated average reward under the (IC) condition and the single CR requirement.

We explain the decision variables and the constraints of the MILP in brief next.

Decision Variables:

- m_{ij}, $(i, j) \in [n] \times [m]$: These binary decision variables correspond to the matching topology.
- p_{ij}, $(i, j) \in [n] \times [m]$: These decision variables approximate the matching probabilities on edge (i, j) under the matching topology $\{m_{ij}\}$ and FCFS-ALIS matching.
- $\eta_i, (i \in [n]); \omega_j, (j \in [m]); \nu_{ij}, ((i, j) \in [n] \times [m])$: These decision variables correspond to the dual variables for flow balance constraints (26a), and non-negativity constraint for p_{ij}, respectively, and are used to enforce the KKT conditions for the quadratic program (**QP**). Recall that the QP dictates that for some constants $\{\eta_i\}_{i\in[n]}$, $\{\gamma_j\}_{j\in[m]}$, and $\{\nu_{ij}\}_{(i,j)\in[n]\times[m]} \geq 0$, we have $p^*_{ij} = \mu_j(\theta_i + \gamma_j + \nu_{ij})$ and $f_{ij} \cdot \nu_{ij} = 0$ if $m_{ij} = 1$, and $f^*_{ij} = 0$ otherwise.

[‡] This MILP is an extension of the one studied in Afèche et al. (2021).

Objective:

$$\mathcal{V}^* := \max \sum_{ij} A_i \cdot p_{ij} \cdot V_{ij}$$

Constraints:

Approximate FCFS matching rates: KKT conditions of the (QP)

$$\sum_{j\in[m]} p_{ij} = 1, \quad \sum_{i\in[n]} f_{ij} = \mu_j, \quad p_{ij} \le Z m_{ij}, \quad p_{ij} \le Z z_{ij}, \quad \nu_{ij} \le (n+m+1)Y \cdot (1 - z_{ij}), \tag{26a}$$

$$\mu_j(\theta_i + \gamma_j + \nu_{ij}) - Z(1 - m_{ij}) \le p_{ij} \le \mu_j(\theta_i + \gamma_j + \nu_{ij}) + Z(1 - m_{ij}), \tag{26b}$$

$$\text{where} \quad Y := \frac{1}{2} \max\left\{\frac{1}{A_{\min}}, \frac{1}{\mu_{\min}}\right\} \quad \text{and} \quad Z := A_{\max} \cdot \mu_{\max} \cdot \left(\frac{n}{A_{\min}} + \frac{m}{\mu_{\min}} + (n+m+1)^2 Y\right)$$

$$A_{\max} = \max_{i\in[n]}\{A_i\}, \quad A_{\min} = \min_{i\in[n]}\{A_i\}, \quad \mu_{\max} = \max_{j\in[m]}\{\mu_j\}, \quad \mu_{\min} = \min_{j\in[m]}\{\mu_j\}.$$

Enforcing incentive compatibility condition (IC):

$$\sum_{j\in[m]} V_{ij}\left(p_{kj} - p_{ij}\right) \le 0. \tag{27}$$

Enforcing a single CRP component: n sets of constraints (indexed by $k \in \mathcal{C}$)

$$\sum_{i\in\mathcal{C}} g_{ij}^{(k)} = \mu_j, \qquad \sum_{j\in\mathcal{S}} g_{ij}^{(k)} = A_i - \frac{\varepsilon}{n-1}, \qquad g_{kj}^{(k)} = A_k + \varepsilon \qquad g_{ij}^{(k)} \le Z m_{i,j}, \tag{28}$$

where $\varepsilon := \left(\prod_{i\in[n]} q_i \prod_{j\in[m]} q_j\right)^{-1}$ and $\frac{p_i}{q_i} = \tilde{A}_i$, $\frac{p_j}{q_j} = \mu_j$ are the rational number representations.

Non-negativity of decision variables: $\{p_{ij}\}, \{\nu_{ij}\}, \{g_{ij}^{(k)}\} \ge 0$ and $\{m_{ij}\}, \{z_{ij}\} \in \{0,1\}$.

Figure 11: MILP for finding a tailored menu with maximum reward rate under minimum average delay constraint.

- z_{ij}, $(i,j) \in [n] \times [m]$: The binary variable $z_{i,j}$ is used to enforce complementary slackness for the non-negativity constraint $p_{ij} \ge 0$: $z_{i,j} = 0$ enforces $p_{ij} = 0$ and $z_{i,j} = 1$ enforces $\nu_{i,j} = 0$.
- $g_{ij}^{(k)}, (i,j,k) \in [n] \times [m] \times [n]$: These n sets of flow variables (where the set is indexed by the superscript $k \in [n]$) are used to enforce the single CRP (equivalently minimum average delay) requirement. In words, the k^{th} set of variables corresponds to the adjusted flows when we increase A_k by a small ε, and reduce each A_i for $i \ne k$ by $\frac{\varepsilon}{n-1}$.

Constraints (26a)-(26b) are the flow balance constraints. The constants Y, Z ensure that the constraints impose the KKT conditions of (**QP**) for any matching topology M. These constraints and the non-negativity of p_{ij} imply:

$$p_{ij} = \begin{cases} \mu_j(\theta_i + \gamma_j + \nu_{ij}), & m_{ij} = 1, \\ 0, & m_{ij} = 0. \end{cases}$$

Constraints (26a) and non-negativity of p_{ij} and ν_{ij} imply the complementary slackness constraint $f_{ij} \cdot \nu_{ij} = 0$. Constraints (27) ensures the IC condition $\overline{V}_{ii} \ge \overline{V}_{ik}$ for all $i, k \in [n]$. Finally, the proof that the constraints (28) are necessary and sufficient to ensure that the matching topology $M = [m_{ij}]$ induces a single CRP component can be found in Afèche et al. (2021).

8 Numerical Experiments

In Sections 6 and 7, we presented two classes of service menus that the service provider can use to design the service system, partition menus and tailored menus. In this section, we perform some preliminary numerical experiments to explore for which preference structures the different approaches perform better or worse.

We begin by showing plots of the performance of Pareto efficient partition menus and tailored menus in the reward-delay quadrant for three particular instances of parameters. The customer valuations for servers were generated using the distribution $V_{ij} = \theta * j + N(0, \sigma)$, where θ and j take values in [5]. Valuations are then translated so that $\min_{ij} V_{ij} = 0$, and scaled so that $\max_{ij} V_{ij} = 10$. We show results for $\sigma = 0, 2$, and 5, $\Gamma = [1,1,1,1,1]$, $A = 5/18[2,5,1,6,4]$, and $a = 1/5[1,1,1,1,1]$. For comparison, we also show a bound on the performance of any menu using a linear programming relaxation of the problem that assumes that the service provider is able to decide the delays and matching rates for each customer type separately, and only the incentive compatibility constraints need to be satisfied. Details of the LP bound can be found in Appendix C. These plots provide some intuition about the relative performance of each menu. We will later show that this intuition applies quite generally, and does not depend on the particular valuations used to generate these plots.

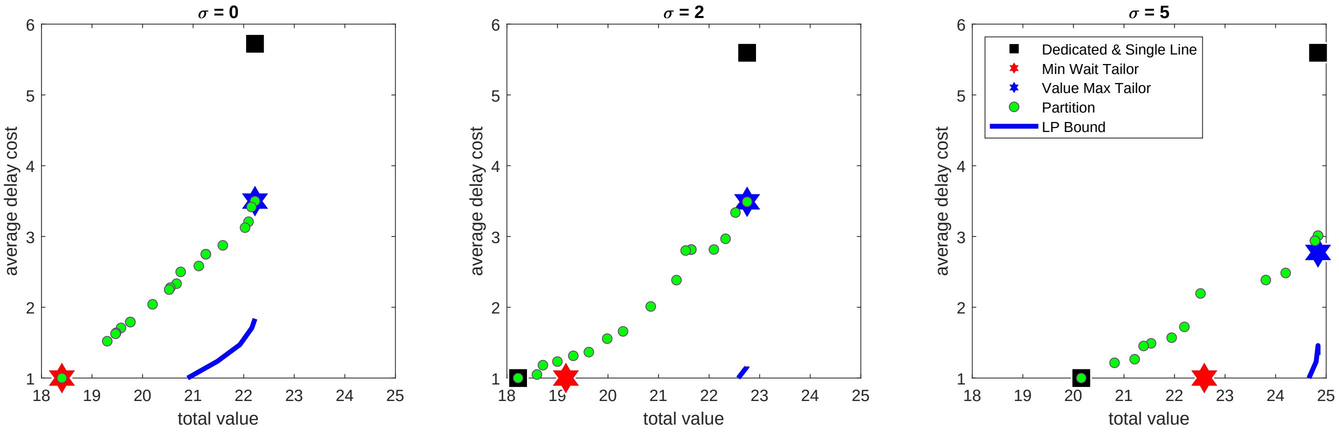

Figure 12: Performance of different menus when $V_{\theta j} = \theta \cdot j + N(0, \sigma)$ for $\sigma = 0, 1, 5$ in the average reward vs. average delay quadrant.

As we can see, the partition menus perform better relative to the tailored menus when there is less noise. The delay minimising tailored menu performs better as the noise increases. The value maximising tailored menu only performs better than some partition menus when noise is large.

Next, we compare the performance of the different mathematical programming approaches for different values of ζ. Using the same valuation distributions and the same values of Γ, A, and a as we used to generate Figure 12, we compare the performance of the chained partition menus and the tailored menus. For $\sigma = 1$ and $\sigma = 5$, we randomly generate 100 different instances of valuations, and report the average performance across all the instances. For $\sigma = 0$, since there is no randomness, we only have one instance, and so we report the performance of that instance directly. For each value of ζ

and each valuation instance, we find the optimal partition menu. The tailored menus do not change depending on ζ, only the objective function values do.

For each menu, we report the ratio

$$\frac{\widehat{V}^* - \zeta\widehat{W}^*}{\widehat{V}_{\text{LP}} - \zeta\widehat{W}_{\text{LP}}}.$$

That is, we are comparing the performance of each menu with that of the LP bound.

σ	0				2				5			
ζ	0	0.05	0.25	0.5	0	0.05	0.25	0.5	0	0.05	0.25	0.5
CP	1.000	0.972	0.864	0.806	1.000	0.986	0.938	0.911	1.000	0.986	0.939	0.904
VM	1.000	0.972	0.851	0.683	1.000	0.986	0.909	0.773	1.000	0.987	0.919	0.812
DM	0.828	0.826	0.818	0.806	0.909	0.919	0.947	0.964	0.957	0.963	0.976	0.982

Table 1: Average performance of different menus when $V_{\theta j} = \theta \cdot j + N(0, \sigma)$ for $\sigma = 0, 1, 5$ relative to the LP bound.

These results show that the intuitions from Figure 12 hold true across many instances, as well as providing some new intuition. For low values of ζ, the optimal partition menu performs at least as well the tailored menus, and when there is no noise, the partition menu performs at least as well as the tailored menus for all values of ζ. For high values of ζ, the delay minimizing tailored menu performs at least as well as the partition menus and the value maximising tailored menus, regardless of how much noise there is. The performance of the delay minimising tailored menu also improves relative to the LP bound as the noise increases. The value maximising tailored menu only outperforms the partition menus when noise is large, and ζ is small.

9 Future Directions and Open Questions

In this work, we have taken the first steps towards studying the design of service systems with congestion in the presence of strategic customers. A key message of our results is that more is not always better – restricting customer choice is as important as offering richer service classes. On the constructive side, we presented a mathematical programming approach to menu design. Our experimental results demonstrate that menus with one service class per type is sufficient to find good menus. In particular, there exist menus which achieve minimum average delay, and at the same time achieve matching value quite close to the optimal. Such menus are appealing for two reasons (i) their simplicity, and (ii) the ability to search within this space through a math programming approach.

Several challenging problems remain towards building a full theory of service menu design; we mention a few. First, we saw empirical evidence that menus with one service class per customer type are sufficient to approximate the Pareto frontier (for two very different reward structures) but lack theoretical bounds. Second, we need a better characterization of the effect of reward structure on the trade-off between matching value and delay on the Pareto frontier. An even simpler question is the following: Given a reward matrix, what is the minimum loss in matching value necessary under a single CRP constraint? This question is quite similar in spirit to the notion of *price of envy-freeness* in the literature on envy-free cake cutting. Again, our experiments indicates this to be small, but it is possible to construct extreme examples where the matching value under a single CRP constraint can be an arbitrarily small fraction of the optimal which makes our experimental results even more intriguing. A third question is on non-uniqueness of equilibrium. We avoided equilibrium selection problem via the notion of provider-preferred equilibrium, but menus with unique equilibria may offer practical advantages such as robustness. Fourth, our results rely on the quasilinear structure of the utility function and homogeneous delay costs. With heterogeneous delay costs, the value optimality of the dedicated menu also breaks down. Extension to more general reward structures, or better yet, menus which are robust to misspecification of utility functions is also an important and challenging direction. Finally, the vast literature on design of price/lead-time menus relies on the *achievable region method* queueing systems where the service provider has full flexibility to dynamically route customers. A similar tool for FCFS-ALIS queueing systems could further expand the menu design settings to which we can apply a mathematical programming approach.

References

I. Adan and G. Weiss. Exact FCFS matching rates for two infinite multitytpe sequences. *Operations Research*, 60(2):475–489, 2012. 3

I. Adan and G. Weiss. A skill based parallel service system under FCFS-ALIS – steady state, overloads and abandonments. *Stochastic Systems*, 4(1):250–299, 2014. 3, 6, 8, 11, 13, 50, 51

I. Adan, A. Bušić, J. Mairesse, and G. Weiss. Reversibility and further properties of FCFS infinite bipartite matching. *Mathematics of Operations Research*, 43(2):598–621, 2018a. 3

I. Adan, I. Kleiner, R. Righter, and G. Weiss. FCFS parallel service systems and matching models. *Performance Evaluation*, 127-128:253–272, 2018b. 3

I. Adan, M. Boon, and G. Weiss. Design heuristic for parallel many server systems. *European Journal of Operations Research*, 273(1):259–277, 2019. 3

P. Afèche. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management*, 15(3):423–443, 2013. 4

P. Afèche and J.M. Pavlin. Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science*, 62(8):2412–2436, 2016. 4

P. Afèche, R. Caldentey, and V. Gupta. On the optimal design of a bipartite matching queueing system. *Operations Research*, 2021. 3, 4, 8, 9, 11, 12, 13, 14, 31, 32

M. Akan, O. Alagoz, B. Ata, F.S. Erenay, and A. Said. A broader view of the liver allocation system incorporating disease evolution. *Operations Research*, 60(4):757–770, 2012. 4

M. Akbarpour, S. Li, and S. Oveis Gharan. Thickness and information in dynamic matching markets. 2018. 4

R. Anderson, I. Ashlagi, D. Gamarnik, and Y. Kanoria. Efficient dynamic barter exchange. *Operations Research*, 65(6):1446–1459, 2017. 4

N. Arnosti and P. Shi. Design of lotteries and waitlists for affordable housing allocation. 2018. 4

N. Arnosti, R. Johari, and Y. Kanoria. Managing congestion in matching markets. 2018. 4

I. Ashlagi, F. Monachou, and A. Nikzad. Optimal dynamic allocation: Simplicity through information design. https://ssrn.com/abstract=3610386, 2021. 4

I. Ashlagi, J. Leshno, P. Qian, and A. Saberi. Price discovery in waiting lists: A connection to stochastic gradient descent. Technical report, University of Chicago, 2022. 4

Itai Ashlagi, Maximilien Burq, Patrick Jaillet, and Vahideh Manshadi. On matching and thickness in heterogeneous dynamic markets. *Operations Research*, 67(4):927–949, 2019. 4

R. Atar. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, 15(4):2606–2650, 2005. 4

M. Baccara, A. Collard-Wexler, L. Felli, and L. Yariv. Child-adoption matching: Preferences for gender and race. *American Economic Journal: Applied Economics*, 6(6):133–158, 2014. 4

Mariagiovanna Baccara, SangMok Lee, and Leeat Yariv. Optimal dynamic matching. *Theoretical Economics*, 15(3):1221–1278, 2020. 4

A. Bassamboo, R.S. Randhawa, and J.A. Van Mieghem. A little flexibility is all you need: On the asymptotic value of flexible capacity in parallel queuing systems. *Operations Research*, 60(6):1423–1435, 2012. 5

S.L. Bell and R.J. Williams. Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electron. J. Probab.*, 10: 1044–1115, 2005. 4

D. Bertsimas, V.F. Farias, and N. Trichakis. Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, 61(1):73–87, 2013. 4

F. Bloch and D. Cantala. Dynamic assignment of objects to queuing agents. *American Economic Journal: Microeconomics*, 9(1):88–122, 2017. 4

A. Bušić, V. Gupta, and J. Mairesse. Stability of the bipartite matching model. *Advances in Applied Probability*, 45(2):351–378, 2013. 3

R. Caldentey and E.H. Kaplan. A heavy traffic approximation for queues with restricted customer-service matchings. Unpublished manuscript, 2002. 3

R. Caldentey, E.H. Kaplan, and G. Weiss. FCFS infinite bipartite matching of severs and customers. *Advances on Applied Probability*, 41(3):695–730, 2009. 3, 13

R. Caldentey, V. Gupta, and L. Hillas. Heavy traffic analysis of multi-class bipartite queueing systems under fcfs. Technical report, University of Chicago, 2022. 9, 10, 12, 20

Francisco Castro, Hamid Nazerzadeh, and Chiwei Yan. Matching queues with reneging: a product form solution. *Queueing Systems*, 96(3):359–385, 2020. 3

J.G. Dai and T. Tezcan. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems*, 59:95–134, 2005. 4

Y. Ding, T. McCormick, and M. Nagarajan. A fluid model for an overloaded bipartite queueing system with heterogeneous matching utility. 2018. 4

M.M Fazel-Zarandi and E.H. Kaplan. Approximating the first-come, first-served stochastic matching model with Ohm's law. *Operations Research*, 6:1423–1432, 2018. 3, 14

K. Gardner and R. Righter. Product forms for fcfs queueing models with arbitrary server-job compatibilities: An overview. *Queueing Systems*, 96:3–51, 2020. 3

L. Green. A queueing system with general-use and limited-use servers. *Operations Research*, 33(1): 168–185, 1985. 3

I. Gurvich and A. Ward. On the dynamic control of matching queues. *Stochastic Systems*, 4(2): 479–523, 2014. 4

I. Gurvich and W. Whitt. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management*, 11(2):237–253, 2009. 4

I. Gurvich and W. Whitt. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research*, 58(2):316–328, 2010. 4

J. M. Harrison and M. J. Lopez. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems Theory Appl.*, 33:339–368, 1999. 4

J.M. Harrison. Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies. *The Annals of Applied Probability*, 8(3):822–848, 1998. 4

W. C. Jordan and S. C. Graves. Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4):577–594, 1995. 5

E.H. Kaplan. Managing demand for publich housing. 1984. ORC Technical Report # 183, MIT. 3

E.H. Kaplan. A public housing queue with reneging and task-specific servers. *Decision science*, 19: 383–391, 1988. 3

J.D. Leshno. Dynamic matching in overloaded waiting lists. 2017. 4

C. Maglaras and A. Zeevi. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research*, 53(2):242–262, 2005. 4

A. Mandelbaum and S. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ rule. *Operations Research*, 52(6):836–855, 2004. 4

R.T. Marler and J.S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, April 2004. ISSN 1615-147X, 1615-1488. doi: 10.1007/s00158-003-0368-6. URL http://link.springer.com/10.1007/s00158-003-0368-6. 29

Mohammadreza Nazari and Alexander L Stolyar. Reward maximization in general dynamic matching systems. *Queueing Systems*, 91(1):143–170, 2019. 4

H. Nazerzadeh and R.S. Randhawa. Near-optimality of coarse service grades for customer differentiation in queueing systems. *Production and Operations Management*, 27(23):578–595, 2018. 4

E. Plambeck. Optimal leadtime differentiation via diffusion approximations. *Operations Research*, 52 (2):213–228, 2004. 4

R. Rogerson, R. Shimer, and R. Wright. Search-theoretic models of the labor market: A survey. *Journal of Economic Literature*, 43(4):959–988, 2005. 4

B.L. Schwartz. Queueing models with lane selection: A new class of problems. *Operations Research*, 22(2):331–339, 2004. 3

Cong Shi, Yehua Wei, and Yuan Zhong. Process flexibility for multiperiod production systems. *Operations Research*, 67(5):1300–1320, 2019. 5

V.M. Slaugh, M. Akan, O. Kesten, and M. Utku Ünver. The pennsylvania adoption exchange improves its matching process. *Interfaces*, 46(2):133–158, 2016. 4

R. Talreja and W. Whitt. Fluid models for overloaded multi-class many-service queueing systems with fcfs routing. *Management Science*, 54(1):1513–1527, 2008. 3

J. N. Tsitsiklis and K. Xu. On the power of (even a little) resource pooling. *Stochastic Systems*, 2(1): 1–66, 2012. 5

J. N. Tsitsiklis and K. Xu. Flexible queueing architectures. *Operations Research*, 65(5):1398–1413, 2017. 5

M.U. Unver. Dynamic kidney exchange. *Rev. Econom. Stud.*, 77(1):372–414, 2010. 4

J.A. Van Mieghem. Price and service discrimination in queuing systems: Incentive compatibility of gcμ scheduling. *Management Science*, 46(9):1249–1267, 2000. 4

R.B. Wallace and W. Whitt. A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management*, 7(4):276–294, 2005. 4

A.R. Ward and M. Armony. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research*, 61(1):228–243, 2013. 4

S.A. Zenios, G.M. Chertow, and L.M. Wein. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research*, 48(4):549–569, 2000. 4

Appendix A: Proofs

Proof of Theorem 1: We will use Kakutani's Fixed Point Theorem to show that a Δ-equilibrium exists for $\Delta = 0$ when $|\alpha| < |\mu|$.

Theorem 8 (Kakutani's Fixed Point Theorem)**.** *Let Q be a non-empty, compact and convex subset of some Euclidean space $\mathbb{R}^n$. Let $F : Q \to 2^Q$ be a set-valued funtcion on X with the following properties:*

- *F has a closed graph;*
- *$F(q)$ is non-empty and convex for all $q \in Q$.*

Then F has a fixed point.

We will apply Kakutani's Fixed Point Theorem to a best response function $F : Q \to 2^Q$, which we will now construct. In constructing the best response function F, it will be useful to extend our definitions of $W_i(q)$, $p_{ij}(q)$, and $U_{\theta i}(q)$ to strategy profiles q for which the system does not admit a steady state under a FCFS-ALIS service discipline.

To do this, we introduce the concept of a reduced service system. Fix a strategy profile q, and let $\lambda_I(q)$ denote the arrival rate into the set I of service classes under q. The strategy profile q need not admit a steady state distribution. Thus, we define $\overline{\mathscr{I}} \subseteq [n]$ as the minimal set of unstable service classes and let $\overline{\mathscr{J}} = S(\overline{\mathscr{I}})$ be the servers compatible with $\overline{\mathscr{I}}$ under menu M. That is, $\overline{\mathscr{I}}$ is the minimal set satisfying $\forall \mathscr{I}' \cap \overline{\mathscr{I}} = \emptyset$:

$$\lambda_{\mathscr{I}'}(q) < \mu_{S(\mathscr{I}') \cap \mathscr{J}},$$

where $\mathscr{J} := [m] \setminus \overline{\mathscr{J}}$, and $\mathscr{I} := [n] \setminus \overline{\mathscr{I}}$. The set $\overline{\mathscr{I}}$ is unique, and a non-constructive method of identification is as follows: Let f^* denote the optimal value of the maximum flow in the network with nodes $[m] \cup [n]$, maximum inflow into service node i of $\lambda_i(q)$ and maximum capacity of server node j of μ_j. If for a service class i, there exists some $\epsilon_i > 0$, such that the new maximum flow obtained by increasing the inflow into service class i by ϵ_i is $f^* + \epsilon_i$ then $i \in \mathscr{I}$, otherwise $i \in \overline{\mathscr{I}}$.

The reduced service system is given by only keeping the service classes $\mathscr{I}$ and servers $\mathscr{J}$. The menu $M^{\mathscr{I}}$ is the submatrix of M with rows corresponding to service classes in $\mathscr{I}$, and columns corresponding to $\mathscr{J}$. We use $\lambda^{\mathscr{I}}(q)$ to denote the vector of arrival rates for service classes $\mathscr{I}$, and $\Gamma^{\mathscr{J}}$ as the service rate vector for the servers in $\mathscr{J}$. Note that the sets $\mathscr{I}$ and $\mathscr{J}$ are a function of the strategy profile q. We will denote them by $\mathscr{I}(q)$ and $\mathscr{J}(q)$ when this dependence is not clear from the context.

We now use this reduced system to define $p_{ij}(q)$ and $W_i(q)$ for arbitrary strategy profiles q (potentially for which the system is unstable), and the best response map. By definition, the reduced service system $(\lambda^{\mathscr{I}}(q), \Gamma^{\mathscr{J}}, M^{\mathscr{I}})$ is stable, and hence admits steady state mean waiting times which we denote by $W_i^{\mathscr{I}}(q)$ for $i \in \mathscr{I}$, and matching probabilities, defined to be $p_{ij}^{\mathscr{I}}(q)$ for $i \in \mathscr{I}, j \in \mathscr{J}$. For $i \in \mathscr{I}, j \in \mathscr{J}$, we set $p_{ij}(q) = p_{ij}^{\mathscr{I}}(q)$. For all other combinations of $(i, j) \in [n] \times [m]$, we set $p_{ij}(q) = 0$. Similarly for $i \in \mathscr{I}$ we set $W_i(q) = W_i^{\mathscr{I}}(q)$, and for all $i \notin \mathscr{I}$ we set $W_i(q) = \infty$.

With these extended definitions of $p_{ij}(q)$ and $W_i(q)$, we can also extend the definition of $U_{\theta i}(q)$ which allows us to define the best response set of each customer type for any strategy profile q. Let $B_\theta(q)$ be the set of all service classes which maximize the utility of customers in class θ given strategy profile q, that is,

$$B_\theta(q) = \left\{ i \in [n] \;\middle|\; i \in \operatorname*{argmax}_{i'} U_{\theta i'}(q) \right\}.$$

For any customer type θ and any strategy profile q, let

$$F_\theta(q) = \operatorname{conv}(\{e_i | i \in B_\theta(q)\}).$$

We then define the best response function $F(q)$ as

$$F(q) = \mathop{\times}_{\theta \in \Theta} F_\theta(q). \tag{A1}$$

It is clear from the definition of F that $F(q)$ is non-empty and convex for all q. All that remains to be shown in order to use Kakutani's Fixed Point Theorem is that the graph of F is closed. To do this, we will show that the graph of F contains all of its limit points.

Let $\{q_k\}_{k\in\mathbb{N}}$ and $\{q_k^*\}_{k\in\mathbb{N}}$ be a sequences of strategy profiles such that $q_k \to q$ and $q_k^* \to q^*$, where q and q^* are strategy profiles, and $q_k^* \in F(q_k)$ for all $k \in \mathbb{N}$. To show that the graph of F contains all of its limits points, we need to show that $q^* \in F(q)$. To do this, we need to show that for all $\theta \in \Theta$ and $i \in [n]$ such that $q_{\theta i}^* > 0$, $q_{\theta i}^* \in B_\theta(q)$.

Consider any pair (θ, i) such that $q_{\theta i}^* > 0$. Then there must exist some $K \in \mathbb{N}$ such that for all $k > K$, $q_{k\theta i}^* > 0$. This implies that $i \in B_\theta(q_k)$ for all $k > K$, or, $U_{\theta i}(q_k) \geq U_{\theta i'}(q_k)$ for all $i' \in [n]$. To show that $U_{\theta i}(q) \geq U_{\theta i'}(q)$ for all $i' \in [n]$, it suffices to show that $U_{\theta i}(q_k) \to U_{\theta i}(q)$ for all θ, i as $k \to \infty$.

Let $\mathscr{I}(q)$ denote the set of stable service classes under limiting strategy profile q. Since $q_k \to q$ implies $\lambda_I(q_k) \to \lambda_I(q)$ for all subsets $I \subseteq [n]$, it is true that $\mathscr{I}(q) = \liminf_{k\to\infty} \mathscr{I}(q_k)$. Further, for all $i \notin \mathscr{I}(q)$ (the unstable service classes), $W_i(q_k) \to \infty$ and hence $\lim_{k\to\infty} U_{\theta i}(q_k) = -\infty = U_{\theta i}(q)$ for $i \notin \mathscr{I}$. For the remaining classes, $i \in \mathscr{I}(q)$, there exists some K, such that for all $k \geq K$, $i \in \mathscr{I}(q_k)$. Thus by continuity of the steady-state distribution for FCFS-ALIS model (for stable matching topologies), for $i \in \mathscr{I}(q)$, $p_{ij}(q_k) \to p_{ij}(q)$ and $W_i(q_k) \to W_i(q)$, and hence $U_{\theta i}(q_k) \to U_{\theta i}(q)$.

This completes the proof that the graph of F is closed. So Kakutani's Fixed Point Theorem applies, and we know that there exists some strategy profile q satisfying $q \in F(q)$. □

Proof of Corollary 1: Note from (6) that

$$w_{\sigma,k} := \sum_{\kappa=\sigma^{-1}(k)}^{K} \frac{1}{\sum_{\ell=1}^{\kappa} \widetilde{\gamma}_{\sigma(\ell)}} = \frac{1}{|a|} + \sum_{\kappa=\sigma^{-1}(k)}^{K-1} \frac{1}{\sum_{\ell=1}^{\kappa} \widetilde{\gamma}_{\sigma(\ell)}}.$$

Let us prove that $w_{\sigma,k} \geq 1/|a|$. From the previous equation, this would follow if the last summation is nonnegative. Suppose, by contradiction that this is not the case. Then, there exists a κ such that $\sigma^{-1}(k) \leq \kappa \leq K-1$ such that $\sum_{\ell=1}^{\kappa} \widetilde{\gamma}_{\sigma(\ell)} < 0$. In other words, the cumulative capacity slack of the CRP components $\{\mathbb{C}_{\sigma(1)}, \mathbb{C}_{\sigma(2)}, \ldots, \mathbb{C}_{\sigma(\kappa)}\}$ is negative. However, this would imply that the cumulative arrival rate to these components exceeds the total service capacity of all the servers in these components. This, together with the DAG structure connecting all the CRP components imply that the stability condition in Proposition 1 is violated, which holds by assumption. From this contradiction we conclude that $w_{\sigma,k} \geq 1/|a|$ and then from (7) we also get that $W_{\mathbb{C}_k} \geq 1/|a|$.

Let us now prove the second part of the corollary, namely, there can be at most one CRP component $\hat{\kappa} \in [K]$ such that $\widehat{W}_{\mathbb{C}_{\hat{\kappa}}} = 1/|a|$. From the previous discussion, it follows that the requirement $\widehat{W}_{\mathbb{C}_{\hat{\kappa}}} = 1/|a|$ can only be satisfied if $w_{\sigma,\hat{\kappa}} = 1/|a|$ for all permutations σ associated a topological order.

But this can only happen if $\sigma^{-1}(\hat{\kappa}) = K$ for all permutation σ. Evidently, this condition can only be satisfied by at most one CRP component and holds trivially if $K = 1$. □

Proof of Proposition 2: Without loss of generality let us index the CRP components in such a way that $W_k = W_{(k)}$ for all $k \in [K]$. We partition the set $[K]$ into equivalence classes $\{\mathscr{C}_1, \ldots, \mathscr{C}_L\}$ such that $i, k \in \mathscr{C}_\ell$ if and only if $W_i = W_k$. We denote by $\mathbb{W}_\ell$ the waiting time of class $\mathscr{C}_\ell$ and by $n_\ell := |\mathscr{C}_\ell|$ its cardinality. We also order these equivalence classes in such that $\mathbb{W}_1 < \mathbb{W}_2 < \cdots < \mathbb{W}_L$. Note that by assumption W is such that $n_1 = 1$.

Next, we show how to implement W using a chained DAG. Define this chained DAG using the partition $\{\mathscr{C}_1, \ldots, \mathscr{C}_L\}$. This is a DAG in which there is a directed arc between $\mathbb{C}_i$ and $\mathbb{C}_k$ if and only if $i \in \mathscr{C}_\ell$ and $k \in \mathscr{C}_{\ell+1}$. Fix a vector of capacity slacks $\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_K)$ that satisfies $\widetilde{\gamma} = |a|$ and $\widetilde{\gamma}_k = \widehat{\gamma}_\ell$ for all $k \in \mathscr{C}_\ell$. It follows from this construction of the DAG and $\tilde{\gamma}$ that for any permutation $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(K))$ induced by some topological order the vector $(\widetilde{\gamma}_{\sigma^{-1}(1)}, \widetilde{\gamma}_{\sigma^{-1}(2)}, \ldots, \widetilde{\gamma}_{\sigma^{-1}(K)})$ is constant. This observation together with Theorem 2 imply that $\mathbb{Q}(\sigma)$ in (6) is also constant, independent of σ. Furthermore, by symmetry it is not hard to see that two CRP components that belong to the same partition $\mathscr{C}_\ell$ have the same limiting scaled waiting times, which we denote by $\widehat{\mathbb{W}}_\ell$. One can show from Theorem 2 that

$$\widehat{\mathbb{W}}_\ell = \widehat{\mathbb{W}}_{\ell-1} + \frac{1}{n_\ell} \sum_{s=1}^{n_\ell} \frac{1}{\sum_{j=\ell+1}^{L} n_j \, \widehat{\gamma}_j + s \, \widehat{\gamma}_\ell}, \qquad \ell = 1, 2 \ldots, L \tag{A2}$$

with $\widehat{\mathbb{W}}_0 = 0$. We use this condition to find the values of $\{\widehat{\gamma}_\ell\}$ that implement $\{\mathbb{W}_\ell\}$, that is, $\widehat{\mathbb{W}}_\ell = \mathbb{W}_\ell$ for all $\ell \in [L]$. To this end, we use backward induction on ℓ. For $\ell = L$ we have that

$$\widehat{\mathbb{W}}_L = \widehat{\mathbb{W}}_{L-1} + \frac{1}{n_L} \sum_{s=1}^{n_L} \frac{1}{s \, \widehat{\gamma}_L}.$$

Thus, $\widehat{\gamma}_L$ must satisfy

$$\widehat{\gamma}_L = \frac{1}{(\mathbb{W}_L - \mathbb{W}_{L-1})} \frac{1}{n_L} \sum_{s=1}^{n_L} \frac{1}{s}.$$

Now suppose that we have determined the values of $\widehat{\gamma}_L, \widehat{\gamma}_{L-1}, \ldots, \widehat{\gamma}_{\ell+1}$ and define $\widehat{\Gamma}_\ell := \sum_{j=\ell+1}^{L} n_j \, \widehat{\gamma}_j$. We find the value $\widehat{\gamma}_\ell$ by solving (A2)

$$\mathbb{W}_\ell = \mathbb{W}_{\ell-1} + \frac{1}{n_\ell} \sum_{s=1}^{n_\ell} \frac{1}{\widehat{\Gamma}_\ell + s \, \widehat{\gamma}_\ell}.$$

We note that there exists a unique $\widehat{\gamma}_\ell$ that solves this equation in the region $\widehat{\gamma}_\ell > -\widehat{\Gamma}_\ell / n_\ell$. This follows from the fact that the summation above is monotonically decreasing in $\widehat{\gamma}_\ell$ in this region and diverges to $+\infty$ as $\widehat{\gamma}_\ell$ approaches $\widehat{\Gamma}_\ell / n_\ell$ from above and converges to zero as $\widehat{\gamma}_\ell$ approaches ∞.

It only remains to show that the vector $\{\widehat{\gamma}_\ell\}$ that we have constructed satisfies the cumulative capacity slack constraint $|\widehat{\gamma}| = |a|$. Since $n_1 = 1$ by assumption, (A2) reduces to

$$\mathbb{W}_1 = \frac{1}{\widehat{\Gamma}_1 + \widehat{\gamma}_1},$$

where the denominator $\widehat{\Gamma}_1 + \widehat{\gamma}_1$ equals $|\widehat{\gamma}|$. The proof is completed by noticing that, by assumption, $\mathbb{W}_1 = W_{(1)} = 1/|a|$. □

Proof of Proposition 4: Since $\hat{q}^*$ is a heavy traffic equilibrium, there exists a direction $\hat{\phi}^* \in \mathbb{R}^{|\Theta|}$ satisfying the conditions in Definition 7. For $\epsilon > 0$, let us define the strategy $q^{(\epsilon)} = \hat{q}^* + \hat{\phi}^* \epsilon$. To prove the result, we will show that $q^{(\epsilon)}$ satisfies condition (a) in the proposition for an appropriate sequence $(\Delta_\epsilon)_{\epsilon>0}$ that converges to 0 as $\epsilon \downarrow 0$. Specifically, we need to show that for all $\theta \in \Theta$ and for all $i, k \in [n]$

$$q_{\theta i}^{(\epsilon)} \left(U_{\theta i}\big(W^{(\epsilon)^{(\epsilon)}}(q^{(\epsilon)}), p^{(\epsilon)}(q^{(\epsilon)})\big) - U_{\theta k}\big(W^{(\epsilon)^{(\epsilon)}}(q^{(\epsilon)}), p^{(\epsilon)}(q^{(\epsilon)})\big) \right) \geq -\Delta^{(\epsilon)}. \tag{A3}$$

Since $\hat{q}^*$ is a heavy traffic equilibrium converging along the direction $\hat{\phi}^*$, conditions (a) and (b) in Definition 7 imply that the left-hand side of this inequality converges to a non-negative limit as $\epsilon \downarrow 0$. It follows then that for all $\Delta > 0$ there exists an $\varepsilon(\Delta) > 0$ such that for all $\epsilon \in (0, \varepsilon(\Delta))$ we have

$$q_{\theta i}^{(\epsilon)} \left(U_{\theta i}\big(\widehat{W}^{(\epsilon)}(q^{(\epsilon)}), p^{(\epsilon)}(q^{(\epsilon)})\big) - U_{\theta k}\big(\widehat{W}^{(\epsilon)}(q^{(\epsilon)}), p^{(\epsilon)}(q^{(\epsilon)})\big) \right) \geq -\Delta.$$

Furthermore, we can always select the mapping $\varepsilon(\Delta) > 0$ to be continuous and monotonically increasing in a neighborhood $(0, \bar{\Delta})$, for some $\bar{\Delta} > 0$, and such that $\lim_{\Delta \downarrow 0} \varepsilon(\Delta) = 0$. Then, for ϵ small enough we can define $\Delta^{(\epsilon)} := \varepsilon^{-1}(\epsilon/2)$. It follows that $\lim_{\epsilon \downarrow 0} \Delta^{(\epsilon)} = 0$ and that the inequality in (A3) is satisfied. □

Proof of Proposition 5: Let us use a slight abuse of notation and denote by $\widehat{W}_k^*$ the limiting scaled waiting times of all service classes that belong to $\mathbb{C}_k$ for $k \in [K]$. Define the vector $\widehat{W}'$ such that $\widehat{W}'_k = \widehat{W}_k^* - \widehat{W}_1^* + 1/|a|$. We next show that $(\hat{q}^*, \widehat{W}', \hat{p}^*)$ is a heavy traffic equilibrium that weakly Pareto dominates $(\hat{q}^*, \widehat{W}^*, \hat{p}^*)$ since $\widehat{W}' \leq \widehat{W}^*$. To this end, we note that by Proposition 2 $\widehat{W}'$ is implementable by a chained DAG on $\mathbb{C}$. Let $\tilde{\gamma}' = (\tilde{\gamma}'_1, \ldots, \tilde{\gamma}'_K)$ be the vector of cumulative capacity slacks that implement $\widehat{W}'$(see Definition 5) and define the direction of convergence to heavy traffic ϕ' as a solution to system of linear equations $\tilde{\gamma}' = a\,\hat{q}^* - A\,\phi'$. By this construction, one case see that $(\hat{q}^*, \widehat{W}', \hat{p}^*)$ satisfies conditions (a) and (b) in Definition 7. Indeed, (a) holds since the sequence of pre-limit strategy profiles $q'^{(\epsilon)} = \hat{q}^* + \epsilon\,\phi'$ satisfies $\widehat{W}' = \lim_{\epsilon \downarrow 0} \widehat{W}^{(\epsilon)}(q'^{(\epsilon)})$ and $\hat{p}^* = \lim_{\epsilon \downarrow 0} p^{(\epsilon)}(q'^{(\epsilon)})$. On the other hand, (b) holds trivially since $\widehat{W}'_k$ is a translation of $\widehat{W}^*$. □

Proof of Proposition 6 : To compute the average scaled waiting time $\overline{W}_{\max}$ under the Dedicated menu we need to impose the equilibrium conditions. First, we need to ensure that the limiting arrival rate to class i converges (from below) to μ_i for $i = 1, 2$. Thus, under the assumption $A_1 > \mu_1$, we must have some customer type $\bar{\theta} \in \Theta_1$ that is randomizing between joining the dedicated queue for server 1 and the dedicated queue for server 2. In equilibrium, this randomization strategy should be such that this customer type is indeed indifferent between joining these two service classes. To identify the type $\bar{\theta}$ we need to rank the customers' types in Θ_1 according to the value of ΔV_θ. For this, let $K_1 = |\Theta_1|$ denote the cardinality of $|\Theta_1|$ and let us index its elements $\Theta_1 = \{\theta_1, \theta_2, \ldots, \theta_{K_1}\}$ in such a way that $\Delta V_{\theta_1} \geq \Delta V_{\theta_2} \geq \cdots \geq \Delta V_{\theta_{K_1}}$. In case of ties, i.e., if $\Delta V_{\theta_i} = \Delta V_{\theta_{i+1}}$, then we require $V_{\theta_i 1} \geq V_{\theta_{i+1} 1}$.

Let us denote by $\bar{\kappa}$ the index that defines $\bar{\theta}$, that is, $\bar{\theta} = \theta_{\bar{\kappa}}$. Now, if type-$\theta_{\bar{\kappa}}$ customers are indifferent between the two service classes then we must have that customers' type θ_k (with $k < \bar{\kappa}$) prefer class 1 over class 2. Hence, to ensure that the arrival rate to class i converges to μ_i from below the value of $\bar{\kappa}$ must be equal to

$$\bar{\kappa} := \min \left\{ \kappa \in [K_1] \colon \sum_{k=1}^{\kappa} A_{\theta_k} > \mu(1) \right\}$$

and a fraction

$$\hat{q}_{\bar{\kappa}} := \frac{\mu_1 - \sum_{k=1}^{\bar{\kappa}-1} A_{\theta_k}}{\sum_{k=1}^{\bar{\kappa}} A_{\theta_k}}$$

of the type-$\theta_{\bar{\kappa}}$ customers must select class 1.

Also, a type-$\theta_{\bar{\kappa}}$ is indifferent between the two dedicated queues if $\Delta V_{\theta_{\bar{\kappa}}} = \delta\,(\widehat{W}_1^{\mathrm{D}} - \widehat{W}_2^{\mathrm{D}})$, where $\widehat{W}_i^{\mathrm{D}}$ is the scaled steady-state mean waiting time of class $i = 1, 2$ under the Dedicated menu. Furthermore, from Theorem 2, we know that $\widehat{W}_i^{\mathrm{D}} = 1/\tilde{\gamma}_i$, where $\tilde{\gamma}_i$ is the scaled capacity slack of class i. But the sum of the slacks of the two classes is equal to the aggregated system lack, that is, $\tilde{\gamma}_1 + \tilde{\gamma}_2 = |a|$. Using this identity, the indifference condition $\Delta V_{\theta_{\bar{\kappa}}} = \delta\,(\widehat{W}_1^{\mathrm{D}} - \widehat{W}_2^{\mathrm{D}})$ leads to the following equation on $\tilde{\gamma}_1$:

$$\Delta V_{\theta_{\bar{\kappa}}} = \delta \left(\frac{1}{\tilde{\gamma}_1} - \frac{1}{|a| - \tilde{\gamma}_1} \right).$$

Solving for $\tilde{\gamma}_1$ and plugging back the solution in the values for $\widehat{W}_1^{\mathrm{D}}$ and $\widehat{W}_2^{\mathrm{D}}$ we get

$$\widehat{W}_1^{\mathrm{D}} = \frac{2\,\Delta V_{\theta_{\bar{\kappa}}}}{2\,\delta + |a|\,\Delta V_{\theta_{\bar{\kappa}}} - \sqrt{4\,\delta^2 + \left(|a|\,\Delta V_{\theta_{\bar{\kappa}}}\right)^2}} \quad \text{and} \quad \widehat{W}_2^{\mathrm{D}} = \frac{2\,\Delta V_{\theta_{\bar{\kappa}}}}{|a|\,\Delta V_{\theta_{\bar{\kappa}}} - 2\,\delta + \sqrt{4\,\delta^2 + \left(|a|\,\Delta V_{\theta_{\bar{\kappa}}}\right)^2}}.$$

Finally, to obtain the value of $\overline{W}_{\max}$ we note that in a Dedicated menu a flow of μ_i customers join class i in the heavy traffic limit, $i = 1, 2$. It follows that

$$\overline{W}_{\max} = \left(\frac{\mu_1}{\mu_1 + \mu_2} \right) \widehat{W}_1^{\mathrm{D}} + \left(\frac{\mu_2}{\mu_1 + \mu_2} \right) \widehat{W}_2^{\mathrm{D}}. \tag{A4}$$

Let us turn to the derivation of $\overline{W}_{\mathrm{med}}$, the average customers' delay performance achieved by the Full and N_1 menus. We can compute this value using a similar line arguments as the one we just used to compute $\overline{W}_{\max}$. Again the equilibrium conditions imply that customers type-$\theta_{\bar{\kappa}}$ (same as above) must randomize between joining class 1 or class 3 and the randomization probability must equal $\hat{q}_{\theta_{\bar{\kappa}}}$ to ensure that the arrival rate to class i converges to μ_i from below in the heavy traffic limit for $i = 1, 3$. The main difference with the Dedicated menu is than under the N_1 menu the two CRP components are not longer disconnected but rather chained. Thus, Theorem 2 implies that the scaled waiting time of class 3 is equal to $\widehat{W}_3^{\mathrm{N}_1} = 1/|a|$. In addition, a type-$\theta_{\bar{\kappa}}$ is indifferent between the two classes if $\Delta V_{\theta_{\bar{\kappa}}} = \delta\,(\widehat{W}_1^{\mathrm{N}_1} - \widehat{W}_3^{\mathrm{N}_1})$ and so $\widehat{W}_1^{\mathrm{N}_1} = 1/|a| + \Delta V_{\theta_{\bar{\kappa}}}/\delta$. Combining these values with the fact that a flow of μ_1 customers join class 1 in equilibrium we get that

$$\overline{W}_{\mathrm{med}} = \frac{1}{|a|} + \left(\frac{\mu_1}{\mu_1 + \mu_2} \right) \frac{\Delta V_{\theta_{\bar{\kappa}}}}{\delta}. \tag{A5}$$

□

Proof of Theorem 3: First, under a Single Line menu every customer –irrespective of its type– is served by server j with probability $\mu_j/|\mu|$. It follows that,

$$\overline{V}^{\mathrm{SL}} = \sum_{\theta \in \Theta} \frac{A_\theta}{|A|} \sum_{j \in [m]} \frac{\mu_j}{|\mu|} V_{\theta j}.$$

On the other hand, let $\widehat{W}^* = (\widehat{W}_i^*)_{i\in[n]}$ and $\hat{p}^* = [\hat{p}^*_{ij}]_{i\in[n],j\in[m]}$ be the limiting steady-state waiting times and matching probabilities under the pair $(M, \hat{q}^*)$. It follows that

$$\overline{V}(M, \hat{q}^*) = \sum_{\theta\in\Theta} \frac{A_\theta}{|A|} \sum_{i\in[n]} \hat{q}^*_{\theta i} \sum_{j\in[m]} \hat{p}^*_{ij}\, V_{\theta j}.$$

Now, let $\hat{\lambda}^*_i$ be the equilibrium arrival rate to class $i \in [n]$ under $(M, \hat{q}^*)$, that is,

$$\hat{\lambda}^*_i = \sum_{\theta\in\Theta} A_\theta\, \hat{q}^*_{\theta i}.$$

and let us define the strategy $q = [q_{\theta i}]_{\theta\in\Theta, i\in[n]}$ by

$$q_{\theta i} = \frac{\hat{\lambda}^*_i}{|\mu|}.$$

Note that q is feasible strategy (i.e., $q \in \mathcal{Q}$) since $|\hat{\lambda}^*| = |A| = |\mu|$.

By the equilibrium condition that $\hat{q}^*$ satisfies, there is no customer type θ that would strictly prefer to use strategy $(q_{\theta i})_{i\in[n]}$ instead of $(\hat{q}^*_{\theta i})_{i\in[n]}$. It follows that

$$\begin{aligned}
\sum_{i\in[n]} \hat{q}^*_{\theta i}\Big(\sum_{j\in[m]} \hat{p}^*_{ij}\, V_{\theta j} - \delta\, \widehat{W}_i\Big) &\geq \sum_{i\in[n]} q_{\theta i}\Big(\sum_{j\in[m]} \hat{p}^*_{ij}\, V_{\theta j} - \delta\, \widehat{W}_i\Big) && \text{(equilibrium condition)}\\
&= \sum_{i\in[n]} \frac{\hat{\lambda}^*_i}{|A|}\Big(\sum_{j\in[m]} \hat{p}^*_{ij}\, V_{\theta j} - \delta\, \widehat{W}_i\Big) && \text{(definition of } \hat{\lambda}^*_i\text{)}\\
&= \sum_{j\in[m]} \frac{V_{\theta j}}{|\mu|} \sum_{i\in[n]} \hat{\lambda}^*_i\, \hat{p}^*_{ij} - \delta \sum_{i\in[n]} \frac{\hat{\lambda}^*_i}{|\mu|}\, \widehat{W}_i\\
&= \sum_{j\in[m]} \frac{\mu_j}{|\mu|}\, V_{\theta j} - \delta \sum_{i\in[n]} \frac{\hat{\lambda}^*_i}{|\mu|}\, \widehat{W}_i. && \text{(since } \textstyle\sum_{i\in[n]} \hat{\lambda}^*_i\, \hat{p}^*_{ij} = \mu_j\text{)}
\end{aligned}$$

Let multiply both sides of the inequality by $A_\theta/|A|$ and sum over $\theta \in \Theta$ to get

$$\overline{V}(M, \hat{q}^*) - \delta \sum_{\theta\in\Theta} \frac{A_\theta}{|A|} \sum_{i\in[n]} \hat{q}^*_{\theta i}\, \widehat{W}_i \geq \overline{V}^{\mathrm{SL}} - \delta \sum_{\theta\in\Theta} \frac{A_\theta}{|A|} \sum_{i\in[n]} \frac{\hat{\lambda}^*_i}{|\mu|}\, \widehat{W}_i.$$

But

$$\sum_{\theta\in\Theta} \frac{A_\theta}{|A|} \sum_{i\in[n]} \hat{q}^*_{\theta i}\, \widehat{W}_i = \sum_{i\in[n]} \frac{\widehat{W}_i}{|A|} \sum_{\theta\in\Theta} A_\theta\, \hat{q}^*_{\theta i} = \sum_{i\in[n]} \frac{\hat{\lambda}^*_i}{|\mu|}\, \widehat{W}_i = \sum_{\theta\in\Theta} \frac{A_\theta}{|A|} \sum_{i\in[n]} \frac{\hat{\lambda}^*_i}{|\mu|}\, \widehat{W}_i \quad \text{(recall that } |A| = |\mu|\text{)}$$

and so we conclude that $\overline{V}(M, \hat{q}^*) \geq \overline{V}^{\mathrm{SL}}$. □

PROOF OF THEOREM 4: We will now use approximate complementary slackness to show that any equilibrium set of flows under the partition menu gives an approximately optimal solution to the max flow problem (**Max-flow**), and that the optimality gap goes to 0 as $\epsilon \downarrow 0$.

Recall that $f^{(\epsilon)}_{\theta j}$ is the flow of customers type θ served by server j in the max flow formulation (**Max-flow**). Similarly, $\eta^{(\epsilon)}_\theta$ and $\omega^{(\epsilon)}_j$ are the dual variables for the flow balance constraint for customer type θ and the capacity constraint of server j, respectively, in the dual problem (**??**). For a feasible primal solution $f_{\theta j}^{(\epsilon)}$ and a feasible dual solution $(\eta_\theta^{(\epsilon)}, \omega_j^{(\epsilon)})$ to satisfy approximate complementary slackness, it is sufficient that

$$\begin{aligned} 0 \quad &\le \left(\mu_j - \textstyle\sum_\theta f_{\theta j}^{(\epsilon)}\right) \omega_j^{(\epsilon)} &&\le \varepsilon_j \\ 0 \quad &\le \left(\eta_\theta^{(\epsilon)} + \omega_j^{(\epsilon)} - V_{\theta j}\right) f_{\theta j}^{(\epsilon)} &&\le \varepsilon_{\theta j} \end{aligned} \tag{A6}$$

for all θ and j, and $\varepsilon_j, \varepsilon_{\theta j} \ll 1$. With these we can show approximate optimality of $f_{\theta j}^{(\epsilon)}$, namely, show that $\sum_{\theta, j} V_{\theta j}\, f_{\theta j}^{(\epsilon)} \approx \bar{\mathcal{V}}^{(\epsilon)}$. Formally, weak duality gives:

$$\sum_{\theta, j} f_{\theta j}^{(\epsilon)}\, V_{\theta j} \le \bar{\mathcal{V}}^{(\epsilon)} \le \sum_\theta \alpha_\theta^{(\epsilon)}\, \eta_\theta^{(\epsilon)} + \sum_j \mu_j\, \omega_j^{(\epsilon)}.$$

Weak complementary slackness and primal/dual feasibility imply:

$$\begin{aligned} \sum_\theta \alpha_\theta^{(\epsilon)}\, \eta_\theta^{(\epsilon)} + \sum_j \mu_j\, \omega_j^{(\epsilon)} &\le \sum_{\theta, j} f_{\theta j}^{(\epsilon)} \eta_\theta^{(\epsilon)} + \sum_j \left(\omega_j^{(\epsilon)} \sum_\theta f_{\theta j}^{(\epsilon)} + \varepsilon_j\right) \\ &= \sum_{\theta, j} f_{\theta j}^{(\epsilon)} \left(\eta_\theta^{(\epsilon)} + \omega_j^{(\epsilon)}\right) + \sum_j \varepsilon_j \\ &\le \sum_{\theta, j} f_{\theta j}^{(\epsilon)}\, V_{\theta j} + \sum_{\theta, j} \varepsilon_{\theta j} + \sum_j \varepsilon_j. \end{aligned}$$

Combining, we get

$$\bar{\mathcal{V}}^{(\epsilon)} - \left(\sum_{\theta, j} \varepsilon_{\theta j} + \sum_j \varepsilon_j\right) \le \sum_{\theta, j} f_{\theta j}^{(\epsilon)}\, V_{\theta j} \le \bar{\mathcal{V}}^{(\epsilon)}. \tag{A7}$$

Let us now show that for any equilibrium arrival rates ${f_{\theta j}}^{(\epsilon)}$[§] for the ϵ^{th} system, we can construct a dual solution such that approximate complementary slackness holds. To this end, let $\widehat{W}_j^{(\epsilon)}$ be the equilibrium limited scaled waiting time for the service class served by server j. For all $j \in [m]$ and $\theta \in \Theta$, we let

$${\omega_j}^{(\epsilon)} = \delta\, \widehat{W}_j^{(\epsilon)} \qquad \text{and} \qquad {\eta_\theta}^{(\epsilon)} = \max_j \left\{V_{\theta j} - {\omega_j}^{(\epsilon)}\right\} \tag{A8}$$

denote a feasible dual solution. We know that under any equilibrium, $f_{\theta k}^{(\epsilon)} > 0$ only ifj

$$j \in \arg\max_{j'} \left\{\overline{V}_{\theta j} - \delta\, \widehat{W}_j^{(\epsilon)}\right\}, \quad \text{that is,} \quad j \in \arg\max_{j'} \left\{\overline{V}_{\theta j} - \omega_j^{(\epsilon)}\right\}.$$

Therefore $f_{\theta j}^{(\epsilon)} > 0$ only if ${\eta_\theta}^{(\epsilon)} + {\omega_j}^{(\epsilon)} - V_{\theta j} = 0$. So *exact* complementary slackness holds for the first set of dual constraints: $\varepsilon_{\theta j} = 0$ for all θ, j. Furthermore, for the primal constraints, we use the fact that under the Dedicated menu service class j operates as a single M/M/1 queue with arrival rate $\sum_\theta {f_{\theta j}}^{(\epsilon)}$ and service capacity j. It follows that the (non-scaled) waiting time in service class j equals $W_j^{(\epsilon)} = 1/(\mu_j - \sum_\theta {f_{\theta j}}^{(\epsilon)})$. Thus, since the scaled waiting time in service class j satisfies $\widehat{W}_j^{(\epsilon)} = \epsilon\, W_j^{(\epsilon)}$, which implies

$$0 \le \left(\mu_j - \sum_\theta {f_{\theta j}}^{(\epsilon)}\right) {\omega_j}^{(\epsilon)} = \delta\, \epsilon. \tag{A9}$$

[§] We know that an equilibrium exists from Theorem 1.

Or, approximate complementary slackness holds with $\varepsilon_j = \delta\,\epsilon$. Then (A7) implies

$$\bar{\mathcal{V}}^{(\epsilon)} - \sum_{\theta,j} f_{\theta j}{}^{(\epsilon)} V_{\theta j} \leq \delta\,\epsilon\,m.$$

So as $\epsilon \downarrow 0$, the difference between the value $V^{(\epsilon)} := \sum_{\theta,j} f_{\theta j}{}^{(\epsilon)} V_{\theta j}$ achieved in equilibrium under the Dedicated menu and the upper bound $\bar{\mathcal{V}}^{(\epsilon)}$ converges to zero.

□

Proof of Theorem 5: To get the main idea across, we first assume that for all customer types θ, the rewards $V_{\theta j}$ are distinct. Later in the proof we remove this assumption.

Let the service classes be labeled so that service class j is the dedicated service class for server j. We begin by claiming that in any heavy traffic equilibrium, with limiting probabilities $\hat{p}^*$, for any service class i there can be at most one server j with $\hat{p}^*_{ij} > 0$. Suppose not, and assume that there are $\hat{p}^*_{ij} > 0$ and $\hat{p}^*_{ij'} > 0$ (for simplicity assume there are only two such servers, the proof generalizes easily). Then servers j and j' must be in the same CRP component. Further service classes j and j' must also be in the same CRP component and hence the limiting scaled mean delay of service classes j and j' equal the limiting scaled mean waiting time for service class i. It can happen that the arrival rate into the dedicated service classes j or j' could be zero, however we can still talk about the virtual waiting time of a customer joining these service classes, and the statement would hold for the limiting scaled virtual waiting time.

Now by assumption, at least one of the dedicated service classes j or j' give strictly higher matching value and no higher delay to any customer type joining class i, and therefore this can not be an equilibrium.

The equilibrium matching system therefore looks as follows: the service classes are partitioned into $m+1$ sets $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m$, such that the classes in $\mathcal{C}_0$ have asymptotically negligible demand: $\mathcal{C}_0 = \{i \in [n] | \Lambda_i = 0\}$. For $j \geq 1$, the service classes in $\mathcal{C}_j$ have asymptotically non-negligible flow to only server j: $\hat{p}^*_{ij} = 1$ for $i \in \mathcal{C}_j$. Again, for this outcome to be an equilibrium, we must have the limiting scaled mean waiting for all service classes within each $\mathcal{C}_j$ ($j \geq 1$) to be equal, and therefore in heavy-traffic limit any customer type will be indifferent between any service class within $\mathcal{C}_j$. We will denote by $\widehat{W}^*_j$ the limiting scaled mean waiting time for service classes in $\mathcal{C}_j$ for $j \geq 2$. To summarize, for any service class $i \in \mathcal{C}_j$, the limiting utility obtained by a customer type θ is $U_{\theta i} = V_{\theta j} - \delta \widehat{W}_j$.

Define:

$$f_{\theta j} = \sum_{i \in \mathcal{C}_j} A_\theta \hat{q}^*_{\theta i}. \tag{A10}$$

as the total flow from customer type θ to server j (through service classes in $\mathcal{C}_j$). Denoting the utility of type θ as

$$U_\theta = \max_i U_{\theta i} = \max_j V_{\theta j} - \delta \widehat{W}_j,$$

best response condition gives $f_{\theta j} > 0$ only if $j \in \operatorname{argmax}_{j'} \left\{ V_{\theta j'} - \delta \widehat{W}_{j'} \right\}$ or equivalently,

$$f_{\theta j} \cdot \left(U_\theta - V_{\theta j} + \delta \widehat{W}_j \right) = 0.$$

But these are precisely the complementary slackness conditions for the maximum value flow linear program (**Max-flow**) in the proof of Theorem 4. Since $[f_{\theta j}]$ is a feasible primal solution, the complementary slackness conditions imply that it is also an optimal, and hence value maximizing, flow.

Removing assumption on rewards: To summarize what we have done so far, we showed that we can use the equilibrium to define the flow matrix $[f_{\theta j}]$, customer utilities U_θ, and server delays $\delta \widehat{W}_j$ so that they are feasible primal dual solutions to the maximum value flow problem (**Max-flow**) and satisfy complementary slackness conditions. We now show we can do so without the restriction on rewards.

Fix a heavy-traffic equilibrium $\hat{q}^*$, and the resulting limiting probabilities $\hat{p}^*$. Define $\widehat{W}_i$ as the limiting mean scaled waiting time of the CRP component that service class j belongs to. Recall the definitions:

$$U_{\theta i} = \sum_j \hat{p}^*_{ij} - \delta \widehat{W}_i,$$

and by the best response condition, the customer utility is defined by

$$U_\theta = \max_i U_{\theta i}.$$

Define $\mathcal{S}^*(i) = \{j : \hat{p}^*_{ij} > 0\}$ as the "effective"set of servers for service class i. Let customer type θ join a service class i (that is, $\hat{q}^*_{\theta i} > 0$) so that $U_\theta = U_{\theta i}$. Suppose $|\mathcal{S}^*(i)| \geq 2$ (the case $|\mathcal{S}^*(i) = 1|$ is vacuously true for the argument). Then we must have that for $j, j' \in \mathcal{S}^*(i)$, $\widehat{W}_j = \widehat{W}_{j'}$ since j, j' are in the same CRP component. It must also be the case that for all $j \in \mathcal{S}^*(i)$, $V_{\theta j}$ are equal. If not, then type θ can deviate to the dedicated service class for the server $j \in \mathcal{S}^*(i)$ with highest reward – this strictly improves the reward and does not incur any further delay disutility. Therefore, for all $j \in \mathcal{S}^*(i)$

$$V_{\theta j} - \delta \widehat{W}_j = U_{\theta i} = U_\theta.$$

Define the total flow from customer type θ to server j, $f_{\theta j}$, as:

$$f_{\theta j} = A_\theta \sum_i \hat{q}^*_{\theta i} \hat{p}^*_{ij}.$$

The preceding arguments imply $f_{\theta j} > 0$ only if $U_\theta = V_{\theta j} - \delta \widehat{W}_j$. We thus again find that $[f_{\theta j}]$, U_θ, $\widehat{W}_j$ define feasible primal-dual solution to (**Max-flow**) satisfying complementary slackness, and hence the flow $[f_{\theta j}]$ maximizes the matching reward. □

Proof of Theorem 6: Suppose the service provider is able to achieve a first best outcome by offering the menu $M*$. As there may be multiple equilibria, throughout this proof we will use equilibrium to mean the first best outcome achieving equilibrium. Let $q^*_{\theta i}$ be the equilibrium strategies, and let p^*_{ij} be the equilibrium matching rates. We will also let $f^*\theta j$ be the equilibrium flows between customer types and servers.

Since a first best outcome is achieved, we know that f^*_{ij} constitute an optimal solution to **Max-flow** with $\epsilon = 0$. We will begin by showing that the positive flows from this solution form a connected graph.

Since a first best outcome is achieved, we know that in equilibirum there is a single CRP component, and hence there is a connected graph between service classes and servers. Since every customer type is joining at least one service class, this implies that there is a path between every customer type and every server, which in turn implies that there is a path between every pair of customer types. Hence the flows between customer types and servers also form a connected graph.

Next we will show that no customer type prefers the **Max-flow** matching outcome of any other customer type. Take any two customer types $\hat{\theta}$ and $\tilde{\theta}$. We will show that $\hat{\theta}$ does not prefer that matching outcome of $\tilde{\theta}$. We will let $\hat{V}$ be the matching value that $\hat{\theta}$ achieves from their equilibrium strategy. Since there is a single CRP component, we know that $\hat{V}$ is the value that $\hat{\theta}$ gains from every service class they are joining in equilibrium, and that $\hat{V}$ is at least as large as the value $\hat{\theta}$ would achieve from joining any other service class.

The value $\hat{\theta}$ gains from $\tilde{\theta}$'s **Max-flow** matching outcome is

$$\begin{aligned} V(\hat{\theta},\tilde{\theta}) &= \sum_i q^*_{\tilde{\theta}i} \sum_j p_{ij} V_{\hat{\theta},j} \\ &\leq \sum_i q^*_{\tilde{\theta}i} \hat{V} \\ &= \hat{V}. \end{aligned}$$

Thus $\hat{\theta}$ prefers their own matching outcome to that of any other customer type. This completes the proof. □

Proof of Corollary 2: The proof of this corollary follows the proof of Theorem 4 for the Dedicated menu essentially verbatim by reinterpreting an individual server in the Dedicated menu by a super-server for each of the partitions with a service capacity equals to the sum of the capacities of the servers in the partition. The only small difference in the proof relates to equation (A9). Specifically, since super-server k does not operates exactly as an M/M/1, it is not longer true that the (non-scaled) waiting time $W_k^{(\epsilon)}$ for service class $\mathcal{C}_k$ is equal to $1/(\mu_{\mathcal{S}_k} - \sum_\theta \hat{f}_{\theta k})$. However, we next show that for all $\epsilon \leq \min_j\{\mu_j\}/((m+1)\,|a|)$, we have

$$\Big(\mu_{\mathcal{S}_k} - \sum_\theta \hat{f}_{\theta k}\Big)\, W_k^{(\epsilon)} \leq 2,$$

which suffices to complete the rest of the steps in the proof of Theorem 4.

To this end, we use the fact that service class $\mathcal{C}_k$ is a single-line multi-server queue with arrival rate $\sum_\theta f_{\theta k}^{(\epsilon)}$ and system utilization $\rho_k^{(\epsilon)} := \sum_\theta f_{\theta k}^{(\epsilon)}/\mu_{\mathcal{S}_k}$. Let us denote by m_k the number of servers in $\mathcal{S}_k$ and by $\{\pi_k^{(\epsilon)}(s)\}$ the stationary distribution of the number of customers in service class k (including those in service). The average number of customers in this class satisfies

$$\begin{aligned} L_k^{(\epsilon)} &= \sum_{s=0}^{\infty} s\,\pi_k^{(\epsilon)}(s) \leq \sum_{s=0}^{m_k-1} m_k\,\pi_k^{(\epsilon)}(s) + \sum_{s=m_k}^{\infty} s\,\pi_k^{(\epsilon)}(s) = m_k + \sum_{s=m_k}^{\infty} (s-m_k)\,\pi_k^{(\epsilon)}(s) \\ &= m_k + \sum_{s=0}^{\infty} s\,\pi_k^{(\epsilon)}(m_k+s) = m_k + \sum_{s=0}^{\infty} s\,(\rho_k^{(\epsilon)})^s\,\pi_k^{(\epsilon)}(m_k) = m_k + \frac{\rho_k^{(\epsilon)}}{(1-\rho_k^{(\epsilon)})^2}\,\pi_k^{(\epsilon)}(m_k). \end{aligned}$$

In the second-to-last equality we have used the brith-death property structure of the system, which implies $\pi_k(s) = (\rho_k^{(\epsilon)})^{s-m_k}\,\pi_k^{(\epsilon)}(m_k)$ for all $s \geq m_k$. We also use this fact to get an upper bound on the value of $\pi_k^{(\epsilon)}(m_k)$ as follows:

$$1 = \sum_{s=0}^{\infty} \pi_k^{(\epsilon)}(s) \geq \sum_{s=m_k}^{\infty} \pi_k^{(\epsilon)}(s) = \sum_{s=m_k}^{\infty} (\rho_k^{(\epsilon)})^{s-m_k}\,\pi_k^{(\epsilon)}(m_k) = \frac{\pi_k^{(\epsilon)}(m_k)}{1-\rho_k^{(\epsilon)}} \quad\Longrightarrow\quad \pi_k^{(\epsilon)}(m_k) \leq 1-\rho_k^{(\epsilon)}.$$

Combining this inequality, the inequality for $L_k^{(\epsilon)}$ above and the fact that $W_k^{(\epsilon)} = L_k^{(\epsilon)} / \sum_\theta f_{\theta k}^{(\epsilon)}$ (by Little's law) we get

$$\Big(\mu_{\mathcal{S}_k} - \sum_\theta f_{\theta k}^{(\epsilon)}\Big)\, W_k^{(\epsilon)} \leq \Big(\mu_{\mathcal{S}_k} - \sum_\theta f_{\theta k}^{(\epsilon)}\Big)\, \frac{m_k}{\sum_\theta \hat{f}_{\theta k}} + 1.$$

By stability we must have $\sum_\theta f_{\theta k}{}^{(\epsilon)} \geq |\alpha^{(\epsilon)}| - (|\mu| - \mu_{\mathcal{S}_k}) = \mu_{\mathcal{S}_k} - |a|\,\epsilon$. We use this inequality to upper bound the right-hand side above to get

$$\Big(\mu_{\mathcal{S}_k} - \sum_\theta f_{\theta k}^{(\epsilon)}\Big)\, W_k^{(\epsilon)} \leq \frac{|a|\, m_k\, \epsilon}{\mu_{\mathcal{S}_k} - |a|\,\epsilon} + 1.$$

Finally, it is not hard to check that for $\epsilon \leq \min_j\{\mu_j\}/((m+1)\,|a|)$ the upper bound above is less than or equal to 2. □

Proof of Proposition 8: Let $(\hat{q}^*, \widehat{W}^{\text{PB}}, \hat{p}^*)$ be the heavy traffic equilibrium under the pure partition menu. From Proposition 7 and the assumption $\omega_1 < \omega_2$ we have that $1/|a| < \widehat{W}_1^{\text{PB}} < \widehat{W}_2^{\text{PB}}$. Thus, $(\hat{q}^*, \widehat{W}^{\text{PB}}, \hat{p}^*)$ satisfies the conditions in Proposition 5. It follows that we can construct another heavy traffic equilibrium $(\hat{q}^*, \widehat{W}^{\text{CB}}, \hat{p}^*)$ that (weakly) Pareto dominates $(\hat{q}^*, \widehat{W}^{\text{PB}}, \hat{p}^*)$ by chaining the CRP components in the pure partition menu. Furthermore, from the proof of Proposition 5 we have that $\widehat{W}_1^{\text{CB}} = 1/|a|$ and $\widehat{W}_k^{\text{CB}} = \widehat{W}_k^{\text{PB}} - \widehat{W}_1^{\text{PB}} + 1/|a|$ as required. Finally, it follows trivially that the two heavy traffic equilibria $(\hat{q}^*, \widehat{W}^{\text{PB}}, \hat{p}^*)$ and $(\hat{q}^*, \widehat{W}^{\text{CB}}, \hat{p}^*)$ produce the same matching value $\bar{\mathcal{V}}$ since they have the same limiting strategy profile $\hat{q}^*$ and matching probabilities $\hat{p}^*$. □

Appendix B: Service Menus with Two Servers

In this section we illustrate the model and solution to the service provider's problem in (1) by characterizing optimal service menus for the special case in which the system has two servers (i.e., $m = 2$). In this setting, we are able to obtain a complete solution as a function of the model's parameters, which provides a number of insights that we will use later to analyze the general case with an arbitrary number of severs. The two-server model is also worth studying in its own right as it provides a parsimonious framework that allows for a non-trivial segmentation of service (e.g., high vs. low quality or fast vs. slow service).

With two servers, there are three possible service classes, namely, Class 1 served only by server 1, Class 2 served only by server 2, and Class 3 served by both servers. With these three classes available, the service provider can offer one of the following five admissible service menus (see Figure 1):[†]

- Dedicated menu (D), in which Classes 1 and 2 are offered,
- Single-line menu (SL), in which only service Class 3 is offered,
- Full menu (F), in which all three classes are offered,
- N_i menu, in which Classes i and 3 are both offered, for $i = 1, 2$.

Performance Analysis in Steady State

In order to derive the equilibrium strategies of these menus we first need to characterize their steady-state performance in terms of waiting times and matching probabilities. To this end, let us fix the service menu M. Since the steady-state analysis of the Dedicated and Single Line menus reduce to those of two M/M/1 and one M/M/2 systems, respectively, we will only discuss the cases in which $M \in \{\mathrm{F}, N_1, N_2\}$.

We derive the steady-state performance of an arbitrary strategy profile $q \in \mathcal{Q}(M)$ using the Markov chain representation of the system proposed by Adan and Weiss (2014) and its corresponding stationary distribution. The following result summarizes this derivation, whose statement make use of the following notation $\Lambda := |\lambda|$, $\Gamma := |\mu|$, $\Delta_i := \mu_i - \lambda_i$, for $i = 1, 2$, $\Delta := \Gamma - \Lambda$ and

$$\mathcal{B} := \left[\frac{\Delta + \lambda_3}{\Delta\, \Delta_1\, \Delta_2} + \frac{1}{\Delta_1\, (\Lambda - \lambda_1)} + \frac{1}{\Delta_2\, (\Lambda - \lambda_2)} + \frac{\Lambda + \lambda_3}{\Lambda\, (\Lambda - \lambda_1)\, (\Lambda - \lambda_2)} \right]^{-1}.$$

Proposition 9. (Steady-State Performance) *Suppose $M \in \{\mathrm{F}, N_1, N_2\}$. Let $q \in \mathcal{Q}(M)$ be a fixed customers' strategy profile, which induces a vector of arrival rates $\{\lambda_i\}_{i\in[n]}$ to the service classes. Then, the steady-state probability that a customer joining Class 3 is served by server 1 and server 2 are equal to*

$$p_{31} = \mathcal{B} \left[\frac{1}{\Lambda\, (\Lambda - \lambda_2)} + \frac{1}{\Delta_2\, (\Lambda - \lambda_2)} + \frac{1}{\Delta\, (\Gamma - \lambda_2)} \left(1 + \frac{\mu_1}{\Delta_2} \right) \right] \quad \textit{and} \quad p_{32} = 1 - p_{31}, \tag{B1}$$

[†] We note that it is possible to offer two additional menus each consisting exclusively of service Class i with $i = 1, 2$. However, the menu that offers only Class i is dominated by menu N_i.

respectively. The steady-state waiting times for the three services classes are given by

$$W_1 = W_3 + \frac{\mathcal{B}}{\Delta_1^2}\left[\frac{1}{\Lambda-\lambda_1}+\frac{1}{\Delta}\right], \quad W_2 = W_3 + \frac{\mathcal{B}}{\Delta_2^2}\left[\frac{1}{\Lambda-\lambda_2}+\frac{1}{\Delta}\right] \quad \textit{and} \quad W_3 = \frac{\mathcal{B}\,(\Delta+\lambda_3)}{\Delta^2\,\Delta_1\,\Delta_2}. \tag{B2}$$

Proof of Proposition 9: Let X denote set of states of the Markov chain proposed by Adan and Weiss (2014) with $x \in X$ a generic state of this Markov chain and $\pi(x)$ its steady state probability distribution. The set X is partitioned into the following subsets:

(a) $x = (s_i, n_i, s_j, n_{j3})$: Both servers are busy with server i serving the oldest arrival, with $i = 1, 2$ and $j = 3 - i$. There are $n_i \geq 0$ customers waiting in the queue of Class i and $n_{j3} \geq 0$ customers waiting in the queues of Classes j and 3 combined. The steady-state probability of x is given by

$$\pi(x) = \mathcal{B}\,\frac{\lambda_i^{n_i}(\lambda_1+\lambda_2+\lambda_3)^{n_{j3}}}{\mu_i^{n_i+1}(\mu_1+\mu_2)^{n_{j3}+1}},$$

for some appropriate normalizing constant $\mathcal{B}$.

(b) $x = (s_i, n_i, s_j)$: Server i is busy and server j is idle. There are $n_i \geq 0$ customers waiting in the queue of Class i and the queues of Classes 1 and 3 are necessarily empty. In this case,

$$\pi(x) = \mathcal{B}\,\frac{\lambda_i^{n_i}}{\mu_i^{n_i+1}(\lambda_j+\lambda_3)}.$$

(c) $x = (s_i, s_j)$: Both servers are idle with server i being idle the longest. In this case,

$$\pi(x) = \frac{\mathcal{B}}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_i+\lambda_3)}.$$

The value of $\mathcal{B}$ is obtained by imposing

$$\sum_{x\in X}\pi(x) = 1.$$

To alleviate the notation, let us define $\Lambda := \lambda_1+\lambda_2+\lambda_3$, $\Gamma := \mu_1+\mu_2$, $\Delta_1 := \mu_1-\lambda_1$, $\Delta_2 := \mu_2-\lambda_2$ and $\Delta := \Gamma-\Lambda$. It follows that

$$\mathcal{B} = \left[\frac{\Delta+\lambda_3}{\Delta\,\Delta_1\,\Delta_2}+\frac{1}{\Delta_1\,(\Lambda-\lambda_1)}+\frac{1}{\Delta_2\,(\Lambda-\lambda_2)}+\frac{\Lambda+\lambda_3}{\Lambda\,(\Lambda-\lambda_1)\,(\Lambda-\lambda_2)}\right]^{-1}. \tag{B3}$$

To calculate the matching probabilities, we first calculate the rate of transitions in the Markov chain associated with a customer from service class 3 beginning service with each server. As the problem is symmetric in the servers, we will only go through the calculations to identify the rate of transitions associated with a class 3 customer beginning service with server 1, which we shall label f_{31}.

The FCFS-ALIS service discipline lets us immediately conclude that there are no transitions from states (s_2, s_1) or (s_1, n_1, s_2) that involve a class 3 customer beginning service with server 1. Any arriving class 3 customer will immediately begin service with the server who has been idle longest, which is server 2 in both cases. In the (s_1, n_1, s_2), we can also see that server 1 completing service will not trigger a class 3 customer beginning service with server 1, as the only waiting customers for server 1 to serve are those that are incompatible with server 2 (i.e., class 1 customers). Similar reasoning tells us that the transitions from state (s_1, s_2) and (s_2, n_2, s_1) associated with a class 3 customer beginning service with server 1 are all of those transitions resulting from a class 3 customer arriving, and hence f_{31} includes the terms $\lambda_3\pi(s_1, s_2)$ and $\lambda_3\pi(s_2, n_2, s_1)$. For $n_1 > 0$, there are no transitions from (s_1, n_1, s_2, n_2) that result in a class 3 customer beginning service with server 1, since as soon as a server 1 finishes serving the customer they are currently serving, they will begin serving another waiting class 1 customer. However, for $n_1 = 0$ and $n_2 > 0$, when server 1 finished serving their current customer, they will begin service with a class 3 customer if a class 3 has been waiting the longest out of those compatible with server 1. This will happen if n_2 consists of x class 2 customers, followed by a class 3 customer. Thus f_{31} will include the term $\mu_1 \sum_{n_2=1}^{\infty} \sum_{x=0}^{n_2-1} \frac{\lambda_2^x \lambda_3}{(\lambda_1+\lambda_2+\lambda_3)^{x+1}} \pi(s_1, 0, s_2, n_2)$. We can use similar reasoning to include that the transition rate also includes the term $\mu_1 \sum_{n_1=1}^{\infty} \sum_{n_2=0}^{\infty} \sum_{x=0}^{n_1-1} \frac{\lambda_1^x \lambda_3}{(\lambda_1+\lambda_2+\lambda_3)^{x+1}} \pi(s_2, n_2, s_1, n_1)$. Thus the total rate of transitions involving a class 3 customer beginning service with server 1 is

$$f_{31} = \mathcal{B}\lambda_3 \left[\frac{1}{\Lambda\left(\Lambda - \lambda_2\right)} + \frac{1}{\Delta_2\left(\Lambda - \lambda_2\right)} + \frac{1}{\Delta\left(\Gamma - \lambda_2\right)} \left(1 + \frac{\mu_1}{\Delta_2}\right) \right]. \tag{B4}$$

The probability that a class 3 customer is server by server 1 is $p_{31} = f_{31}/\lambda_3$.

To conclude the proof, we note that the expected waiting times for the different service class can be calculated using Little's Law. □

Equilibrium Strategies

The key feature of the two-server model that we exploit to derive customers' equilibrium strategies is the fact that we can rank the customer types based on their relative preferences over the two servers. To this end, define $\Delta V_\theta := V_{\theta 2} - V_{\theta 1}$ for each customer type $\theta \in [\Theta]$ and label the elements in $[\Theta]$ by $\theta_1, \theta_2, \ldots, \theta_\Theta$ such that $\Delta V_{\theta_i} \leq \Delta V_{\theta_j}$ for all $1 \leq i < j \leq \Theta$. In case of a tie, the class that values server 2 more gets assigned a higher index.

Under this indexing, it is not hard to see that we can restrict ourselves to cut-off (threshold-type) equilibria. For example, if the service provider offers a Dedicated menu then a type θ customer (weakly) prefers Class 1 over Class 2 if $\Delta V_\theta \leq \delta\left(W_2 - W_1\right)$. Thus, there exists a customer type θ_τ with $\tau \in [\Theta]$ such that all customer types θ_k with $k \leq \tau - 1$ select Class 1, all customer types θ_k with $k \geq \tau + 1$ select Class 2 and customers of type θ_τ are indifferent and randomize between the two service classes. Similarly, if the service provider offers the Full menu then a type θ customer weakly prefers Class 1 to Class 3 if $p_{32}\,\Delta V_\theta \leq \delta\left(W_3 - W_1\right)$ and weakly prefers Class 2 to Class 3 if $p_{31}\,\Delta V_\theta \geq \delta\left(W_2 - W_3\right)$. In this case, an equilibrium involves two thresholds, $\tau_1, \tau_2 \in [\Theta]$ with $\tau_1 \leq \tau_2$. All customer types θ_k with $k \leq \tau_1 - 1$ select Class 1, all customer types θ_k with $k \geq \tau_2 + 1$ select Class 2, all customer types θ_k with $\tau_1 + 1 \leq k \leq \tau_2 - 1$ select Class 3, customers type θ_{τ_i} are indifferent between Class i and Class 3 for $i = 1, 2$.

Proposition 10 below exploits this threshold structure to characterize equilibrium strategies for the D, N_1 and F menus[‡]. The statement of this proposition make use of some additional notation. For $0 \leq x_1 \leq x_2 \leq \Theta$, we define

$$\lambda_1(x) := \sum_{k=1}^{\lfloor x \rfloor} \alpha_{\theta_k} + \left(x - \lfloor x \rfloor\right) \alpha_{\theta_{\lceil x \rceil}},$$

[‡] The equilibrium strategy for the Single Line is trivial and for the N_2 menu it can be derived from the one for the N_1 menu by interchanging the labels of the two servers.

$\lambda_2(x_2) = |\alpha| - \lambda_1(x_2)$ and $\lambda_3(x_1, x_2) = |\alpha| - \lambda_1(x_1) - \lambda_2(x_2)$. These are the arrival rates to service classes 1, 2 and 3, respectively, if all customers type $\{1, 2, \ldots, \lfloor x_1 \rfloor\}$ plus a fraction $(x_1 - \lfloor x_1 \rfloor)$ of customers type $\lceil x_1 \rceil$ join Class 1, all customers type $\{\lceil x_2 \rceil + 1, \ldots, \Theta\}$ plus a fraction $(\lceil x_2 \rceil - x_2)$ of customers type $\lceil x_2 \rceil$ join Class 2, and all remaining customers join Class 3. To ensure stability, we will need to bound the values of x_1 and x_2 such that $0 \le x_1 \le \bar{x}_1$ and $\underline{x}_2 \le x_2 \le \Theta$ with

$$\bar{x}_1 := \max\Big\{0 \le x \le \Theta \colon \lambda_1(x) \le \mu_1\Big\} \quad \text{and} \quad \underline{x}_2 := \min\Big\{0 \le x \le \Theta \colon \lambda_2(x) \le \mu_2\Big\}.$$

Note that under the global stability condition $|\alpha| < |\mu|$ we must have $\underline{x}_2 < \bar{x}_1$. For a pair $(x_1, x_2) \in [0, \bar{x}_1) \times (\underline{x}_2, |\Theta|] \cap \{x_1 \le x_2\}$, we define the steady-state matching probabilities $p_{3j}(x_1, x_2)$ and waiting times $W_i(x_1, x_2)$ for $i = 1, 2, 3$ and $j = 1, 2$ by replacing the values $\lambda_1(x_1)$, $\lambda_2(x_2)$ and $\lambda_3(x_1, x_2)$ in equations (B1) and (B2), respectively.

Proposition 10. *Suppose the service provider offers menus $M \in \{D, N_1, F\}$. There exists two thresholds $0 \le x_1^* \le x_2^* \le \vartheta$ such that an equilibrium profile $(q^*_{\theta_k 1}, q^*_{\theta_k 2}, q^*_{\theta_k 3})$ for a type-θ_k customer satisfies*

$$q^*_{\theta_k 1} = \begin{cases} 1 & \text{if } k \le \lceil x_1^* \rceil - 1 \\ x_1^* - \lfloor x_1^* \rfloor & \text{if } k = \lceil x_1^* \rceil \\ 0 & \text{if } k \ge \lceil x_1^* \rceil + 1 \end{cases} \qquad q^*_{\theta_k 2} = \begin{cases} 0 & \text{if } k \le \lceil x_2^* \rceil - 1 \\ \lceil x_2^* \rceil - x_2^* & \text{if } k = \lceil x_2^* \rceil \\ 1 & \text{if } k \ge \lceil x_2^* \rceil + 1 \end{cases}$$

and $q^*_{\theta_k 3} = 1 - q^*_{\theta_k 1} - q^*_{\theta_k 2}$. The values of x_1^* and x_2^* depends on the specific menu M as follows:

–) Dedicated Menu*: Let $x^* = \sup\big\{x \in (\underline{x}_2, \bar{x}_1) \colon \Delta V_{\theta_{\lceil x \rceil}} \le \delta\left(W_2(x, x) - W_1(x, x)\right)\big\}$. If $x^* \notin \mathbb{N}$ then $x_1^* = x_2^* = x^*$. Otherwise, $x_1^* = x^* + 1$ and $x_2^* = x^*$.*

–) N_1 Menu*: $x_1^* = \sup\Big\{x \in [0, \underline{x}_2 \wedge \bar{x}_1) \colon\ p_{32}(x, x_2^*)\, \Delta V_{\theta_{\lceil x \rceil}} \le \delta\left(W_3(x, x_2^*) - W_1(x, x_2^*)\right)\Big\}$ and $x_2^* = \Theta$.*

–) Full Menu*: The values of x_1^* and x_2^* solves the system of equations*

$$\begin{aligned} x_1^* &= \sup\Big\{x \in [0, x_2^* \wedge \bar{x}_1) \colon \quad p_{32}(x, x_2^*)\, \Delta V_{\theta_{\lceil x \rceil}} \le \delta\left(W_3(x, x_2^*) - W_1(x, x_2^*)\right)\Big\} \\ x_2^* &= \sup\Big\{x \in (\underline{x}_2 \vee x_1^*, |\Theta|] \colon \quad p_{31}(x_1^*, x)\, \Delta V_{\theta_{\lceil x \rceil}} \le \delta\left(W_2(x_1^*, x) - W_3(x_1^*, x)\right)\Big\}. \end{aligned}$$

An example of the equilibrium strategies derived in Proposition 10 is depicted in Figure 2. While Proposition 10 provides a complete characterization of customers' equilibrium strategies for the Dedicated, N_1 (N_2), and Full menus we can only derive these equilibria computationally for any particular set of parameters. Furthermore, as we try to move to more complex systems with an arbitrary number of servers, we are no longer able to rank customer types based on their preferences over just two servers and use the simple cut-off analysis that we have used above to derive their equilibrium strategies. For this reason, and to say something more concrete about equilibrium outcomes for general systems, we will investigate their performance under heavy traffic conditions.

Proof of Proposition 10: The proof of the proposition follows from noticing that in the equilibrium of each of the three menus some customer type(s) needs to randomize between two service classes to ensure that the equilibrium condition are satisfied. It is easy to see that the values of x_i^*, $i = 1, 2$ specify precisely the customer types that need to randomized. □

Appendix C: Numerics

Here we include the LP used to find an upper bound under FCFS-ALIS scheduling on the performance of any menu used in section Section 8.

The following are the decision variables used in the LP. formulation:

-) $p_{\theta j}$: probability that customer type θ is served by server j.
-) $f_{\theta j}$: flow of type-θ customers to server j.
-) W_θ: waiting time for type θ customers.

OBJECTIVE:

$$\sum_{\theta\in\Theta} A_\theta \sum_{j\in[m]} p_{\theta j} V_{\theta j} - \zeta \sum_{\theta\in\Theta} A_\theta W_\theta \tag{B5}$$

CONSTRAINTS:

Flow balance:

$$\sum_j p_{\theta j} = 1, \qquad \sum_\theta A_\theta p_{\theta j} = \mu, \qquad f_{\theta j} = A_\theta p_{\theta j}. \tag{B6}$$

Waiting time constraint:

$$W_\theta \geq \frac{1}{\sum_\theta a_\theta} \tag{B7}$$

Incentive compatibility:

$$\sum_j \left(p_{\theta j} - p_{\theta' j}\right) V_{\theta j} + W_{\theta'} - W_\theta \geq 0. \tag{B8}$$

Non-negativity of decision variables:

$$\{p_{\theta j}\}, \{f_{\theta j}\}\{W_\theta\} \geq 0. \tag{B9}$$

Figure 13: LP for finding an upperbound on the performance of any menu.