Service Rate Differentiation for Homogeneous Impatient Customers

Chenguang (Allen) Wu and Wei You

Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong SAR, China, allenwu@ust.hk, weiyou@ust.hk

Abstract

We study joint service rate and waiting time differentiation for homogeneous impatient customers in many-server quality-based queueing systems, where longer services generate higher values but customers may abandon if they wait too long. Customers are homogeneous upon arrival, but the system manager can differentiate them in two dimensions: by assigning each customer to one of many service grades corresponding to different service rates, and by using customers' elapsed waiting times as a scheduling signal to further differentiate those who are waiting in queue. The manager jointly chooses the service grades, an allocation rule that assigns arriving customers to grades, and a scheduling policy that prioritizes customers across and within each grade. Because an exact stochastic analysis of this joint optimization problem is intractable, we adopt a fluid approach and translate the fluid solution into implementable policies for the underlying stochastic system. Our main structural result is that there exists an optimal fluid policy with at most two active service rate and offered wait pairs, regardless of the welfare function or patience time distribution. This reduces the complicated design problem to operationally simple policies. Motivated by our fluid analysis, we propose using a randomized grade assignment at customer arrival and a priority rule based on grade labels and elapsed waiting times, and establish a formal performance guarantee for this policy in Markovian systems. Our framework can easily extend to heterogeneous customers, where we show at most one customer class needs to be further differentiated, thereby suggesting a robust principle of implementing joint service rate and waiting time differentiation.

Keywords: service rate differentiation, time-in-queue scheduling, queues

1 Introduction

We study the joint optimization of service rate and waiting time differentiation for homogeneous impatient customers in many-server queueing systems, motivated by call centers and other quality-based services such as primary care, legal advisory, diagnostic centers, and premium digital advisory platforms. A defining feature of these settings is that service quality scales with service duration: a short interaction delivers limited value, whereas a longer one can generate substantially higher value. At the same time, customers are impatient and will abandon the system if they wait too long

for service to begin, so longer service times also mean fewer customers served. This tension between service quality and system congestion is empirically documented in call centers (Hu et al. 2022) and is central to a growing literature on customer-intensive services (Hopp et al. 2007, Anand et al. 2011, Xu et al. 2015, Wang et al. 2023).

To navigate this tension, the manager can differentiate otherwise identical customers along two dimensions. First, the *service rate*: each arriving customer is assigned to one of several grades, each associated with a distinct service rate and therefore a distinct service quality. Second, the *waiting time*: customers who are not served upon arrival join a queue, and as they wait, their elapsed waiting times reveal information about their residual patience, which can be exploited through scheduling.

These two dimensions are inherently coupled. Grade differentiation creates the heterogeneity that waiting time differentiation exploits, and waiting time differentiation is in turn what allows the manager to reap the full benefits of grade differentiation. Taken together, they give rise to three coherent decisions: the *service grade design* (the number of grades and the service rate of each), an *allocation rule* that assigns arriving customers to grades, and a *scheduling policy* that prioritizes customers both across grades and within each grade using grade labels and elapsed waiting times. Our goal is to jointly optimize these three decisions to maximize the long-run welfare aggregated over all arriving customers.

An exact stochastic analysis of the joint design, allocation, and scheduling problem is intractable. Hence, we adopt a fluid-based approach. We solve the joint problem in a fluid model, characterize the optimal policy, and translate the fluid solution into implementable rules for the underlying stochastic system. This approach yields both structural insights into when and how the two dimensions of differentiation create value, and provides operational guidance on how to jointly implement them.

1.1 Contributions

We summarize our contributions as follows.

- **A unified framework for joint service rate and waiting time differentiation.** To the best of our knowledge, this is the first paper that jointly optimizes service grade design, grade allocation, and scheduling for homogeneous impatient customers in many-server systems. Prior works have treated service rate differentiation (Xu et al. 2015, Wang et al. 2020) and time-in-queue scheduling (Bassamboo and Randhawa 2016, Bassamboo et al. 2023) in isolation. By endogenizing the class structure that downstream scheduling can exploit, our framework bridges these two streams and shows how they can be integrated to create value.

- **A sharp structural characterization: at most two active pairs.** Our central result (Proposition 1) is that, regardless of the welfare function or patience time distribution, there exists an optimal fluid solution with at most two active service rate and offered wait pairs. This reduces a high-dimensional design problem to operationally simple policies: single-grade FCFS, single-grade with two offered waits, or two-grade with FCFS within each grade. This also highlights how customer abandonment and the many-server context can drive different

findings from prior single-server results—such as a continuum of grades (Xu et al. 2015) and exactly two grades (Wang et al. 2020).

- **Reward-shape conditions for when service rate differentiation matters.** Under separable welfare functions, we identify conditions on reward functions that determine when a critically loaded system is optimal and when service rate differentiation is not necessary (Proposition 3). These conditions are independent of the patience time distribution and cover several reward functions commonly used in the quality-based service literature (Anand et al. 2011, Xu et al. 2015, Wang et al. 2023); see Table 1 for a summary.

- **Hazard-rate–driven policy prescriptions.** When waiting costs are material, the hazard rate of the patience time distribution determines the specific format of the optimal policy (Propositions 5, 6, and 7, and summary in Table 2). A particularly interesting result is that two-grade overloaded systems can be optimal under decreasing hazard rates of patience times, and a common offered wait is applied to both grades (Proposition 6). This allows service rate differentiation to occur without waiting time differentiation, extending the results in Bassamboo and Randhawa (2016) and Bassamboo et al. (2023), both focusing on waiting time differentiation exclusively.

- **An implementable $\mathcal{S}$ policy with performance guarantees.** We translate the fluid solution into an implementable stochastic-system policy that uses randomized grade assignment at customer arrival and a priority rule based on customers' grade labels and elapsed waiting times. For Markovian systems, we show this policy attains an $\mathcal{O}(\sqrt{\Lambda})$ optimality gap, and further sharpen the gap to $o(1)$ when a single-grade overloaded system is optimal (Proposition 8). We also develop an equivalent waiting-time-first implementation in Appendix A, showcasing two interchangeable views of the fluid optimization problem.

- **Robustness to heterogeneous customers.** We extend our analysis to heterogeneous customers and show that *at most one customer class needs to be further differentiated*, providing managerial guidance for multi-class environments where fairness or class-specific service level constraints may apply.

- **Numerical evidence on the value of differentiation.** Simulations across exponential, Erlang, hyper-exponential, and log-normal patience time distributions show significant welfare gains from service rate differentiation as large as 90% under homogeneous customers and 55% under heterogeneous customers. We also show that patience distributions with similar means but different hazard rate shapes can prescribe very different optimal policies and system regimes.

1.2 Literature Review

Our work is related to the literature on service rate control in queueing systems. One stream of literature focuses on balancing the service quality, which increases with the service time, with system

congestion, in the context of quality-based or customer-intensive services (Hopp et al. 2007, Anand et al. 2011, Wang et al. 2023). Others focus on the dynamic control of service rates to minimize the cumulative waiting cost and service-effort related cost for single-server queues (George and Harrison 2001, Ata and Shneorson 2006, Adusumilli and Hasenbein 2010, Kumar et al. 2013). Lee and Kulkarni (2014) extends this analysis to multi-server queues. Some studies integrate these two streams of literature by considering service rate differentiation for quality-based services. Xu et al. (2015) and Wang et al. (2020) study the system manager's grade decisions for single-server queues without customer abandonment. The scheduling policy is either a simple $c\mu$-rule in Xu et al. (2015) or not relevant at all in Wang et al. (2020). Armony and Yom-Tov (2026) uses a fluid approach to study the truncation of customers' service times for many-server queues without customer abandonment. The optimal regime in Armony and Yom-Tov (2026) is always critically loaded, so waiting time differentiation on the fluid scale is not critical in their setting. Our paper adds to this literature by considering impatient customers, and demonstrates the highly nontrivial impacts of customers' patience distributions on the optimal grade design and scheduling policy.

Since scheduling is an integral component of our proposed policies, our work is related to a growing literature on scheduling multi-class customers and servers in queueing systems. Among them, Gurvich and Whitt (2010) studies how to minimize the staffing cost for large-scale systems with multiple customer classes and agent pools, subject to service-level constraints. Armony and Mandelbaum (2011) studies how to match homogeneous customers with heterogeneous servers. Atar et al. (2010) and Long et al. (2020) propose the $c\mu/\gamma$-priority policy to minimize the long-run average queue-length cost for systems with multi-class customers. Our paper is considerably different from this literature because we consider customers that are homogeneous at the time of arrival, but the system manager strategically differentiates them in two dimensions: service rates and waiting times in queue.

Time-in-queue scheduling that differentiates customers on waiting times is first proposed in Bassamboo and Randhawa (2016) for single-class systems and extended by Bassamboo et al. (2023) to multi-class systems. A fundamental assumption in these works is that each customer's class identity (corresponding to her service rate) is fixed and exogenous, whereas such decisions are jointly optimized with waiting times in our setting. Despite this important distinction, our problem has intimate connections with these existing works. Our fluid formulation generates grade differentiation upon customers' arrivals and uses this information to implement the multi-class scheduling policy in Bassamboo et al. (2023). In other words, our policy *endogenizes* the class primitives that are necessary inputs for applying the scheduling policy in Bassamboo et al. (2023). In Appendix A, we present an equivalent view of the fluid formulation that first differentiates customers based on their elapsed waiting times and then creates service rate differentiation among non-abandoning customers when they enter service. Thus, our equivalent formulation builds on the waiting time differentiation (Bassamboo and Randhawa 2016) and extends this approach to enable joint service rate differentiation.

2 Model

2.1 Queueing System

We consider a queueing system in which homogeneous customers[1] arrive at the system according to a stationary renewal process with an arrival rate of Λ. Customers are served by a single pool of N agents, each processing incoming work at a unit rate. Each customer is associated with a patience time and will abandon the system if her service does not start within this time after she arrives. We denote the cumulative distribution function of customers' patience times by $F(\cdot)$, the probability density function by $f(\cdot)$, and the reciprocal of the mean patience time by γ. Upon arrival, each customer is irrevocably assigned to one of K grades. Each grade is specified from a family of service time distributions $\{G_\mu : \mu \in [\underline{\mu}, \overline{\mu}]\}$, where μ and $1/\mu$ are the mean service rate and mean service time of G_μ, respectively. The system manager chooses the service rate of each grade, and that choice determines the associated service time distribution of that grade. We assume $[\underline{\mu}, \bar{\mu}]$ constitutes the feasible set of service rates, with $\underline{\mu} < \mu_0 := \Lambda/N < \bar{\mu}$. The number of grades K and the service rate vector $\boldsymbol{\mu} := (\mu_1, \mu_2, \ldots, \mu_K)$ are to be determined by the system manager.

The goal of the system manager is to maximize the long-run system welfare aggregated over all arriving customers. To define this metric, let $\xi(x, y, w)$ denote the net utility experienced by a customer with service time x, patience time y, and offered wait w, where the offered wait is the amount of time the customer would wait to enter service if she were infinitely patient. Then, for a grade-i customer with service rate μ_i, the average surplus associated with offered wait w is

$$r(\mu_i, w) = \int_0^\infty \int_0^\infty \xi(x, y, w) dG_{\mu_i}(x) dF(y). \tag{1}$$

In the above, the dependence of $r(\mu_i, w)$ on the service time distribution is fully captured by the chosen service rate μ_i through G_{μ_i}; this allows us to write the average surplus as a function of (μ_i, w) only. We assume that $r(\mu, w)$ is decreasing in μ and w. Specifically, customers incur disutility from waiting so that $r(\mu, w)$ is decreasing in w. The service is quality-based, and longer service times can induce higher service quality so that $r(\mu, w)$ is decreasing in μ (Anand et al. 2011, Xu et al. 2015, Wang et al. 2023). Hu et al. (2022) finds further empirical evidence of this latter assumption in call centers. A possible reason is that higher service speeds can lead to increased service failures, which drive customer complaints and adversely affect their service rewards (Zhan and Ward 2019).

As a special case, a commonly used customer utility function in the literature has the following separable form

$$\xi(x, y, w) = U(x)\mathbb{I}(y \geq w) - c \min\{y, w\}, \tag{2}$$

where $U(x)$ represents a customer's reward from receiving a service that lasts x time units, and c represents the customer's cost per unit time waiting in the queue. (2) is separable in the sense that

[1] We extend our analysis to heterogeneous customers that differ in patience times, arrival rates, and welfare functions in §6, and show that our main insight of creating at most one additional active service rate and offered wait pair beyond the m exogenous customer classes carries through.

a customer's waiting cost does not depend on her service time.[2] Plugging (2) into (1), we have

$$r(\mu_i, w) = \bar{F}(w)V(\mu_i) - c\int_0^w \bar{F}(y)dy, \tag{3}$$

where the grade-i reward function

$$V(\mu_i) := \int_0^\infty U(x)dG_{\mu_i}(x) \tag{4}$$

is decreasing in μ_i. Existing literature has considered several special forms of $V(\cdot)$. For example, Anand et al. (2011) and Wang et al. (2023) assume $V(\mu) = R - \alpha\mu$ for $R, \alpha > 0$ and Xu et al. (2015) assumes $V(\mu) = \alpha/\mu$ for $\alpha > 0$.

In this paper, we will first work with a general welfare function (1) to develop our basic results. We will then focus on separable welfare functions (3) and (4) to establish more structures of the optimal policy.

The system manager utilizes joint service rate and waiting time differentiation, corresponding to policies on grade allocation and scheduling, respectively, to maximize the system welfare. In this paper, we consider non-anticipating (non-forward-looking) grade allocation and scheduling policies. Specifically, a grade allocation policy is described by a K-dimensional service rate vector $\boldsymbol{\mu} = (\mu_1 \ldots, \mu_K)$ and probability vector $\boldsymbol{p} = (p_1, \ldots, p_K)$, where $\sum_{i=1}^K p_i = 1$, such that each arriving customer is assigned to grade i with probability p_i (note that the number of grades K is itself a decision variable). A scheduling policy π utilizes the above grade information and specifies how to route idle servers to waiting customers, and it comprises two decisions: when a server becomes available, which customer grade to allocate the server to, and within that grade, which customer in the queue to allocate the server to.

Formally, the system manager aims to find a grade allocation policy $(\boldsymbol{\mu}, \boldsymbol{p})$ and a scheduling policy π to maximize the system welfare aggregated over all arriving customers,

$$\max_{\boldsymbol{\mu},\boldsymbol{p},\pi} \sum_{i=1}^K \Lambda p_i \mathbb{E}[\xi(X_i^\pi, Y, W_i^\pi)], \tag{5}$$

where W_i^π denotes the steady-state offered wait experienced by a representative grade-i customer under policy π. The expectation is taken over the random variables X_i, Y and W_i^π.

2.2 Fluid Model

Noting the analytical intractability of an exact solution, we focus on solving the grade allocation and scheduling optimization problem (5) using a fluid model. We then translate the fluid solution to implementable policies for stochastic systems.

In the fluid model, customers arrive at the system in the form of a fluid deterministically and continuously at the corresponding arrival rate and are processed at deterministic rates based on

[2]One can also consider a waiting cost that includes a customer's time in service, and this can be captured by setting $\xi(x, y, w) = [U(x) - cx]\mathbb{I}(y \geq w) - c\min\{y, w\}$ and $V(\mu_i) := \int_0^\infty U(x)dG_{\mu_i}(x) - c/\mu_i$. Thus, all our results will qualitatively extend to this alternative formulation.

their grade information. Further, the capacity of servers is considered in a fluid manner too, and can be allocated fractionally. We note that there is an intimate relationship between the grade allocation and scheduling policies for the fluid model and the original queueing system. The fluid version of a grade allocation policy specifies the arrival rate to each grade and the rate at which each grade is processed. The fluid version of a scheduling policy has two decisions (Bassamboo and Randhawa 2016): the first decision of which customer grade to allocate an idle server to is equivalent to computing the fraction of capacity allocated to each grade; the second decision of which customer within a grade to allocate the server to can be solved by splitting that grade into multiple subgrades (each representing a unique offered wait) and processing them under FCFS.

Formally, we formulate our grade allocation and scheduling optimization problem in the fluid model in two stages: the first stage focuses on assigning customers to grades, and the second stage optimizes the capacity allocation to each grade and the capacity allocation to each subgrade within the grade. The first-stage *grade allocation* problem, which specifies the number of grades, service rate of each grade, and how to allocate customers to grades, can be stated as

$$\max_{\boldsymbol{\mu},\boldsymbol{p}} \ \mathcal{R}(\boldsymbol{\mu},\boldsymbol{p}), \tag{6}$$

where $\mathcal{R}(\boldsymbol{\mu},\boldsymbol{p})$ is the optimal system welfare under the grade allocation $(\boldsymbol{\mu},\boldsymbol{p})$. Under this allocation, $\mathcal{R}(\boldsymbol{\mu},\boldsymbol{p})$ is achieved by implementing waiting time differentiation both across grades and within each grade, and is attained as the objective value of the following second-stage *scheduling* problem:

$$\mathcal{R}(\boldsymbol{\mu},\boldsymbol{p}) = \max_{\boldsymbol{n}:\sum_i n_i \le N} \sum_i R(n_i|\mu_i,p_i), \tag{7}$$

where $R(n_i|\mu_i,p_i)$ is the optimal welfare of grade i by allocating capacity n_i to that grade.

To characterize the welfare $R(n_i|\mu_i,p_i)$ for grade i, we divide each grade i (with arrival rate Λp_i and service rate μ_i) into multiple subgrades such that each subgrade is processed under FCFS. We use $J(i)$ to denote the number of subgrades created from grade i. (If $J(i)=1$, then there is only one subgrade for grade i and it comprises the entire grade.) We denote the arrival rate to subgrade $j=1,\ldots,J(i)$ of grade i by $\lambda_{i,j}$ and the capacity allocated to this subgrade by $n_{i,j}$. Since the entire grade i has a service rate μ_i, customers of subgrade j are processed deterministically at rate $n_{i,j}\mu_i$ in the fluid model.

Under FCFS, the offered wait for subgrade j of grade i solves a "rate-balance" equation so that the rate of customers entering service (accounting for customer abandonment) equals the total service rate entailed by the processing capacity allocated to that subgrade. In this way, if $\lambda_{i,j} \ge n_{i,j}\mu_i$, the offered wait $w_{i,j}$ solves

$$\lambda_{i,j}\bar{F}(w_{i,j}) = n_{i,j}\mu_i, \tag{8}$$

(we assume that the patience distribution F is strictly increasing, so that (8) above has a unique solution) and otherwise, if $\lambda_{i,j} < n_{i,j}\mu_i$, then $w_{i,j}=0$ because this subgrade has excess capacity to process all arrivals without delays. Thus, the offered wait $w_{i,j}$ solves

$$\lambda_{i,j}\bar{F}(w_{i,j}) = \min\{n_{i,j}\mu_i, \lambda_{i,j}\}. \tag{9}$$

Note that each subgrade j, $j = 1, \ldots, J(i)$ of grade i is characterized by a unique service rate and offered wait pair $(\mu_i, w_{i,j})$. The total welfare $R(n_i|\mu_i, p_i)$ of grade i is obtained by summing up the welfare over all its subgrades and is equal to the objective value of the following within-grade optimization problem:

$$
\begin{aligned}
R(n_i|\mu_i, p_i) = \sup_{\{J(i), n_{i,j}, w_{i,j}, \lambda_{i,j}, j=1,\ldots,J(i)\}} \quad & \sum_j \lambda_{i,j} r(\mu_i, w_{i,j}) \\
\text{s.t.} \quad & \sum_{j=1}^{J(i)} \lambda_{i,j} = \Lambda p_i, \\
& \lambda_{i,j} \bar{F}(w_{i,j}) \le n_{i,j} \mu_i, \\
& \sum_j n_{i,j} \le n_i.
\end{aligned} \tag{10}
$$

The last constraint $\sum_j n_{i,j} \le n_i$ allows us to withhold capacity when necessary.

In this formulation, the grade-allocation problem (6), the cross-grade capacity-allocation problem (7), and the within-grade capacity-allocation problem (10) jointly define the fluid optimization problem for grade allocation and scheduling.

Appendix A presents an equivalent formulation of the fluid optimization problem. In that formulation, customers are first differentiated by their elapsed waiting times, and when they enter service, are assigned (possibly different) service rates depending on their elapsed waiting times. We prove in Appendix A an equivalence between the two formulations. Because they are equivalent, we focus on the first formulation to develop our results of the optimal policy.

2.3 Preliminary Analysis

Under our first policy formulation, homogeneous customers are segregated into subclasses specified by their service rate and offered wait pairs. We present a structural property of the optimal solution below; it specifies the total amount of differentiation needed to achieve optimality.

Proposition 1. *There exists an optimal solution to* (6)*,* (7)*, and* (10) *that creates at most two classes i and i' with different service rate and offered wait pairs, i.e., $(\mu_i, w_i) \neq (\mu_{i'}, w_{i'})$.*

To understand Proposition 1, we reformulate the fluid optimization (6), (7), and (10) as a linear program of fluid allocations with two active constraints (corresponding to total arrival rate and total capacity, respectively). Then, standard linear programming theory suggests that there exists an optimal solution with at most two active fluid allocations. The queueing model enters this fluid optimization problem by generating feasible columns: each service-rate/offered-wait pair (μ, w) yields welfare $r(\mu, w)$ and consumes capacity $\bar{F}(w)/\mu$. The structure then follows from the two constraints.

Remark 1 (Scope of Proposition 1)**.** *The "at most two" structure in Proposition 1 is exact for the fluid model. In finite stochastic systems, an optimal policy need not have the same form. Rather, the fluid result provides an asymptotic benchmark and motivates implementable policies whose*

performance can be evaluated in stochastic systems (see §5). Moreover, additional operational constraints, such as grade-specific capacity caps, fairness constraints, minimum-service requirements, or exogenous limits on the fraction assigned to a premium grade, may also change the number of active pairs. Thus, Proposition 1 identifies the minimal differentiation structure for the simplest fluid model, providing a basis for further extensions.

In view of Proposition 1, it suffices to consider the following three types of policies.

Corollary 1. *There exists an optimal solution to* (6), (7), *and* (10) *such that it corresponds to one of the following three cases:*

1. *it comprises one grade and processes that grade under FCFS.*
2. *it comprises one grade and processes that grade using two different offered waits.*
3. *it comprises two different grades and processes each grade under FCFS.*

Notably, Proposition 1 and Corollary 1 hold under *all* welfare functions and patience time distributions. An important implication is that, although our model allows differentiation in both customers' service rates and waiting times, the optimal policy uses at most two active service rate and offered wait pairs. More specifically, the optimal fluid solution either consists of a single grade that differentiates customers by offered waits, or of two grades, serving each under FCFS. This simple structure makes the policy operationally tractable with minimal deviation from single-grade FCFS and, in the two-grade case, preserves a natural notion of fairness within each grade.

Remark 2 (Comparison with prior service rate differentiation works)**.** *Several prior works study service rate differentiation in single-server systems without customer abandonment, but reach qualitatively different conclusions about the optimal number of grades. Xu et al. (2015) considers a single-server queueing system without customer abandonment and shows that it is optimal to create a continuum of grades to maximize system welfare. Wang et al. (2020) considers another single-server system without customer abandonment wherein customers not receiving service upon arrival will enter an orbit queue and repeatedly revisit the system until ultimately served. Wang et al. (2020) proposes joint service rate and retrial rate (the rate of revisiting the system if not served in prior visits) differentiation to reduce customers' expected waiting times, and shows that the optimal policy creates exactly two grades. Unlike these works, we study many-server systems with customer abandonment and establish that the optimal policy has* at most *two grades. Our analysis in §3.2 and §4 further suggests that all three types of policies in Corollary 1 can emerge as the optimal solution. In particular, there are cases in which the optimal policy consists of* only *one grade.*

Following Corollary 1, we only need to consider two optimization problems formulated in (11) and (13) below (where we equivalently optimize over arrival rates $\boldsymbol{\lambda} := \Lambda \boldsymbol{p}$ instead of allocation probabilities $\boldsymbol{p}$) and select the solution that yields a higher objective value. In both problems, the total capacity N is fully utilized in optimality. (Otherwise, if excess capacity exists, the system will be better off by reducing service rates to fully utilize the excess capacity.)

The first two cases in Corollary 1 combine into the following single-grade optimization problem:

$$\mathcal{M}_1^* = \sup_{\mu\in[\underline{\mu},\bar{\mu}]} \mathcal{M}_1(\mu), \quad \text{where} \quad \mathcal{M}_1(\mu) = \sup_{\lambda_l,\lambda_h,w_l,w_h} \lambda_l r(\mu, w_l) + \lambda_h r(\mu, w_h) \tag{11}$$
$$\text{s.t.} \quad \lambda_l + \lambda_h = \Lambda,$$
$$\lambda_l \bar{F}(w_l) + \lambda_h \bar{F}(w_h) = N\mu, \quad \lambda_l, \lambda_h \geq 0.$$

The first case of Corollary 1 corresponds to $(w_l, w_h) = (w_l, \bar{w})$ or $(\bar{w}, w_h)$ for any $w_l < \bar{w} < w_h$, where $\bar{w}$ solves $\Lambda\bar{F}(\bar{w}) = N\mu$. In general, for (11) to have a feasible solution, it must hold that $\mu \leq \mu_0 = \Lambda/N$. (Otherwise if $\mu > \mu_0$, there is excess capacity even if $w_l = w_h = 0$ and the system can be better off by reducing μ to μ_0 to fully utilize the excess capacity.) We thus focus on $\mu \leq \mu_0$ without loss of generality. Whenever $w_l < \bar{w} < w_h$, feasible arrival rates (λ_l, λ_h) can be uniquely determined by (w_l, w_h) as follows,

$$\lambda_l = \frac{N\mu - \Lambda\bar{F}(w_h)}{\bar{F}(w_l) - \bar{F}(w_h)}, \quad \lambda_h = \Lambda - \lambda_l, \tag{12}$$

so it suffices to optimize over (w_l, w_h). In (11), we rank the offered waits such that $w_l < w_h$.

An important special case of $\mathcal{M}_1(\mu)$ is $\mathcal{M}_1(\mu_0)$. Since $\lambda_l + \lambda_h = \Lambda = N\mu_0$ and $\lambda_l\bar{F}(w_l) + \lambda_h\bar{F}(w_h) = N\mu_0$, any feasible solution at $\mu = \mu_0$ must satisfy $\bar{F}(w) = 1$ for every active subgrade, and thus every active subgrade has offered wait 0. Hence, $\mathcal{M}_1(\mu_0)$ corresponds to a critically loaded (CL) system: in the fluid model all customers enter service immediately. We define $\mathcal{M}_1^{\mathrm{CL}} := \mathcal{M}_1(\mu_0)$.

For $\mu < \mu_0$, any feasible solution has positive fluid abandonment because $\Lambda - N\mu > 0$. Thus, at least one active subgrade must have a positive offered wait. This leads to an overloaded (OL) system and we define $\mathcal{M}_1^{\mathrm{OL}} := \sup_{\mu\in[\underline{\mu},\mu_0)} \mathcal{M}_1(\mu)$. Combining both, we have $\mathcal{M}_1^* = \max\{\mathcal{M}_1^{\mathrm{CL}}, \mathcal{M}_1^{\mathrm{OL}}\}$.

The third case in Corollary 1 can be described by the following two-grade optimization problem:

$$\mathcal{M}_2^* = \sup_{\underline{\mu}\leq\mu_2<\mu_1\leq\bar{\mu}} \mathcal{M}_2(\mu_1, \mu_2), \quad \text{where} \quad \mathcal{M}_2(\mu_1, \mu_2) = \sup_{\lambda_1,\lambda_2,w_1,w_2} \lambda_1 r(\mu_1, w_1) + \lambda_2 r(\mu_2, w_2) \tag{13}$$
$$\text{s.t.} \quad \lambda_1 + \lambda_2 = \Lambda,$$
$$\lambda_1\bar{F}(w_1)/\mu_1 + \lambda_2\bar{F}(w_2)/\mu_2 = N,$$
$$\lambda_1, \lambda_2 \geq 0.$$

Since we assume $\mu_2 < \mu_1$ (we can always do so by indexing the grades properly), for (13) to have a feasible solution, it must hold that $\mu_2 \leq \mu_0$. In contrast, the feasible μ_1 can be either higher or lower than μ_0. Holding (μ_1, μ_2) fixed, whenever $\min\{\frac{\bar{F}(w_1)}{\mu_1}, \frac{\bar{F}(w_2)}{\mu_2}\} \leq \frac{1}{\mu_0} \leq \max\{\frac{\bar{F}(w_1)}{\mu_1}, \frac{\bar{F}(w_2)}{\mu_2}\}$, the feasible arrival rates (λ_1, λ_2) can be uniquely determined by (w_1, w_2) as follows,

$$\lambda_1 = \frac{N - \Lambda\bar{F}(w_2)/\mu_2}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}, \quad \lambda_2 = \Lambda - \lambda_1, \tag{14}$$

so it suffices to optimize over (w_1, w_2). Importantly, we do not impose $w_1 < w_2$ as we do in (11). In fact, the relative values of w_1 and w_2 can be arbitrary.

An important observation of (11) and (13) is that in both cases, once the service rate(s) are determined, one needs to further optimize *two* offered waits to fully solve the joint grade allocation and scheduling problem. Specifically, the cross-grade capacity allocation (7) is irrelevant in the single-grade problem (11) because only one grade is considered; the within-grade capacity allocation (10) is irrelevant in the two-grade problem (13) because the within-grade policy is FCFS. In both cases, we can convert the original fluid optimization problem (6), (7), and (10) to a new two-stage optimization problem that is easier to solve: the first stage focuses on finding the optimal service rate(s), and the second stage optimizes the offered wait(s) for each (sub)grade.

3 Optimizing Fluid Model

In this section, we provide more structure of the optimal solution to the two-stage optimization problems (11) and (13). Recall that the first stage focuses on finding the optimal service rate(s), and the second stage optimizes the offered wait(s) for each (sub)grade. We first characterize necessary conditions for the second-stage optimal offered waits in §3.1, and then focus on separable welfare functions (3) to establish general results of the optimal policy in §3.2.

3.1 Optimizing Offered Waits

Recall from (12) and (14) that, under fixed service rates, the offered waits determine the arrival rates and required capacities on the feasible domain. In the single-grade problem, if $\bar{w}$ solves $\Lambda\bar{F}(\bar{w}) = N\mu$, feasibility requires $w_l \le \bar{w} \le w_h$, and then (λ_l, λ_h) is uniquely determined by (w_l, w_h). In the two-grade problem, feasibility requires $\min\{\frac{\bar{F}(w_1)}{\mu_1}, \frac{\bar{F}(w_2)}{\mu_2}\} \le \frac{1}{\mu_0} \le \max\{\frac{\bar{F}(w_1)}{\mu_1}, \frac{\bar{F}(w_2)}{\mu_2}\}$, and (λ_1, λ_2) is uniquely determined by (w_1, w_2) whenever $\bar{F}(w_1)/\mu_1 \ne \bar{F}(w_2)/\mu_2$. We next present necessary first-order conditions for optimal offered waits.

Proposition 2.

1. *Consider the single-grade optimization problem* (11) *under* $\mu \le \mu_0$. *Let* $\bar{w}$ *solve* $\Lambda\bar{F}(\bar{w}) = N\mu$.

   (i) An optimal solution (w_l, w_h) *such that* $0 < w_l < \bar{w} < w_h < \infty$ *must satisfy*

   $$\frac{\frac{\partial r(\mu, w_l)}{\partial w_l}}{f(w_l)} = \frac{\frac{\partial r(\mu, w_h)}{\partial w_h}}{f(w_h)} = -\frac{r(\mu, w_l) - r(\mu, w_h)}{\bar{F}(w_l) - \bar{F}(w_h)}.$$

   (ii) An optimal solution of the form $(0, w_h)$ *with* $\bar{w} < w_h < \infty$ *must satisfy*

   $$\frac{\frac{\partial r(\mu, w_h)}{\partial w_h}}{f(w_h)} = -\frac{r(\mu, 0) - r(\mu, w_h)}{1 - \bar{F}(w_h)}.$$

   (iii) An optimal solution of the form (w_l, ∞) *such that* $0 < w_l < \bar{w}$ *must satisfy*

   $$\frac{\frac{\partial r(\mu, w_l)}{\partial w_l}}{f(w_l)} = -\frac{r(\mu, w_l) - r(\mu, \infty)}{\bar{F}(w_l)}.$$

2. Consider the two-grade optimization problem (13) *under* (μ_1, μ_2) *such that* $\min\{\mu_1, \mu_2\} \le \mu_0$.

(i) An optimal solution (w_1, w_2) *such that* $0 < w_1, w_2 < \infty$ *must satisfy*

$$\frac{\frac{\partial r(\mu_1, w_1)}{\partial w_1} \cdot \mu_1}{f(w_1)} = \frac{\frac{\partial r(\mu_2, w_2)}{\partial w_2} \cdot \mu_2}{f(w_2)} = -\frac{r(\mu_1, w_1) - r(\mu_2, w_2)}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}.$$

(ii) An optimal solution of the form $(0, w_2)$ *such that* $0 < w_2 < \infty$ *must satisfy*

$$\frac{\frac{\partial r(\mu_2, w_2)}{\partial w_2} \cdot \mu_2}{f(w_2)} = -\frac{r(\mu_1, 0) - r(\mu_2, w_2)}{1/\mu_1 - \bar{F}(w_2)/\mu_2}.$$

The case of $(w_1, 0)$ *with* $0 < w_1 < \infty$ *follows analogously.*

(iii) An optimal solution of the form (w_1, ∞) *such that* $0 < w_1 < \infty$ *must satisfy*

$$\frac{\frac{\partial r(\mu_1, w_1)}{\partial w_1}}{f(w_1)} = -\frac{r(\mu_1, w_1) - r(\mu_2, \infty)}{\bar{F}(w_1)}.$$

The case of (∞, w_2) *with* $0 < w_2 < \infty$ *follows analogously.*

Part 1 and Part 2 provide necessary first-order conditions for the optimal offered waits in the single-grade optimization problem (11) and the two-grade optimization problem (13), respectively. These conditions are obtained by substituting the arrival rates (solved as functions of offered waits) into the welfare objective and differentiating the objective with respect to offered waits. Intuitively, holding the service rate(s) fixed, lowering the offered wait of one (sub)grade increases the welfare of that (sub)grade but consumes more capacity. In contrast, raising the offered wait releases capacity but lowers welfare. When an interior solution (of offered waits) is optimal, the marginal values/costs of adjusting offered waits must balance across all active (sub)grades. This balance is captured in the conditions in Proposition 2.

A general form of the welfare function $r(\mu, w)$ precludes a more refined analysis of the optimal policy. Next, we consider separable welfare functions (3) to obtain more refined results.

3.2 Optimizing Fluid Model under Separable Welfare Functions

In this section we analyze how to optimize (11) and (13) under the separable welfare function (3). We normalize the delay cost c to 1 to simplify our notation without loss of generality. Under (3), for a subgrade j of grade i with service rate μ_i, arrival rate $\lambda_{i,j}$, offered wait $w_{i,j}$, and capacity $n_{i,j}$, its total welfare is

$$\lambda_{i,j} r(\mu_i, w_{i,j}) = \min\{n_{i,j}\mu_i, \lambda_{i,j}\} V(\mu_i) - \lambda_{i,j} \int_0^{w_{i,j}} \bar{F}(y) dy,$$

where $\min\{n_{i,j}\mu_i, \lambda_{i,j}\}$ and $\lambda_{i,j} \int_0^{w_{i,j}} \bar{F}(y) dy$ are the expected throughput and queue-length of this subgrade in the steady state, respectively.

To proceed with our analysis, we define $\mathcal{M}_2^{\mathrm{CL}}$ as follows:

$$\mathcal{M}_2^{\mathrm{CL}} := \sup_{\underline{\mu} \le \mu_2 < \mu_1 \le \bar{\mu}} \mathcal{R}^{\mathrm{CL}}(\mu_1, \mu_2) := \lambda_1 V(\mu_1) + \lambda_2 V(\mu_2) \tag{15}$$
$$\text{s.t.} \quad \lambda_1 + \lambda_2 = \Lambda, \quad \lambda_1/\mu_1 + \lambda_2/\mu_2 = N, \quad \lambda_1, \lambda_2 \ge 0.$$

Namely, $\mathcal{M}_2^{\mathrm{CL}}$ is the *optimal* welfare of a two-grade critically loaded system. Relatedly, recall that $\mathcal{M}_1^{\mathrm{CL}} = \mathcal{M}_1(\mu_0)$ is defined as the welfare (not necessarily optimal) of a single-grade critically loaded system. By definition, it is evident that $\mathcal{M}_2^{\mathrm{CL}} \ge \mathcal{M}_1^{\mathrm{CL}}$. We will later provide cases in which the equality is actually attained and thus the two-grade policy reduces to a single-grade policy.

Although in principle the second-stage optimal offered waits should depend on the patience time distribution (cf. Proposition 2), the results below hold under all patience time distributions.

Proposition 3. *Consider the welfare function* (3).

1. *If $V(\mu) \cdot \mu$ is increasing, then the optimal regime is critically loaded. Thus, the optimal system welfare is $\mathcal{M}_2^{\mathrm{CL}}$.*

2. *If $V(\mu) \cdot \mu$ is concave, then the optimal policy should have only one grade.*

Part 1 considers the optimal queueing regime. If $V(\mu) \cdot \mu$ is increasing, then increasing the service rate(s) can increase the service rewards per unit of capacity while cutting customers' waiting times and queue length. This is so until the system becomes critically loaded, in which case, all fluid customers are served without delays. Further increasing the service rate(s) will not affect the already-zero offered wait or the already-fully-served arrivals, but only reduce the service reward per served customer. Hence, the optimal queueing regime must be critically loaded with system welfare $\mathcal{M}_2^{\mathrm{CL}}$.

Part 2 considers the total number of service grades in optimality. If $V(\mu) \cdot \mu$ is concave, then for any two-grade allocation (μ_1, μ_2) with $\mu_1 \ne \mu_2$, one can find a dominant single-grade allocation with a better system welfare (by properly constructing the single-grade service rate μ as a convex combination of μ_1 and μ_2). Thus, the optimal policy should have only one grade. Moreover, assuming a concave $V(\mu) \cdot \mu$ is not very restrictive, as it is satisfied by several commonly used reward functions in the literature (see Table 1 below).

Remark 3 (Common reward functions in literature)**.** *In Table 1, we summarize common reward functions $V(\cdot)$ in the literature on quality-based services, along with the monotonicity and convexity of $V(\mu) \cdot \mu$. Applying Proposition 3, the optimal policy operates a single-grade critically loaded system under $V(\mu) = V_0 - \alpha\mu^\theta$ (Anand et al. 2011, Wang et al. 2023) and $V(\mu) = \alpha/\mu^\theta$ with $\theta \le 1$ (Xu et al. 2015). Under $V(\mu) = \alpha/\mu^\theta$ with $\theta > 1$, the optimal policy depends on the patience time distribution and we examine such dependence through a more refined analysis in §4.*

The optimal system welfare of a two-grade critically loaded system, $\mathcal{M}_2^{\mathrm{CL}}$, will play a critical role in our subsequent analysis. We thus provide more structure of this important concept. Note that $\mathcal{M}_2^{\mathrm{CL}}$ defined in (15) has $(w_1, w_2) = (0, 0)$, so only service rates are to be optimized.

Table 1: Commonly used reward functions: structures and optimal policies

$V(\mu)$	$V(\mu)\cdot\mu$	Optimal Policy
$V_0-\alpha\mu^\theta$ (large V_0, $\theta>0$)	increasing, concave	single-grade, CL
α/μ^θ $(0<\theta<1)$	increasing, concave	single-grade, CL
α/μ	constant	single-grade, CL
α/μ^θ $(\theta>1)$	decreasing, convex	Depends on Patience Distribution

Lemma 1. *Consider the two-grade optimization problem* (15) *in the critically loaded regime.*

1. *The optimal* (μ_1,μ_2) *such that* $\underline{\mu}<\mu_2<\mu_1<\bar{\mu}$ *must satisfy*

$$-V'(\mu_1)\cdot\mu_1^2=-V'(\mu_2)\cdot\mu_2^2=\frac{V(\mu_1)-V(\mu_2)}{1/\mu_1-1/\mu_2}.$$

2. *If the optimal* (μ_1,μ_2) *has the form of* $(\mu_1,\underline{\mu})$ *with* $\underline{\mu}<\mu_1<\bar{\mu}$*, it must hold that*

$$-V'(\mu_1)\cdot\mu_1^2=\frac{V(\mu_1)-V(\underline{\mu})}{1/\mu_1-1/\underline{\mu}}.$$

3. *If the optimal* (μ_1,μ_2) *has the form of* $(\bar{\mu},\mu_2)$ *with* $\underline{\mu}<\mu_2<\bar{\mu}$*, it must hold that*

$$-V'(\mu_2)\cdot\mu_2^2=\frac{V(\bar{\mu})-V(\mu_2)}{1/\bar{\mu}-1/\mu_2}.$$

To understand Lemma 1, note that in a critically loaded system all customers are served, so the manager is effectively splitting a fixed amount of average workload $1/\mu_0$ across two grades. Let $s=1/\mu$ denote the mean service time. Define $U(s):=V(1/s)$ and then $U'(s)=-V'(\mu)\mu^2$. This represents the marginal gains of welfare from slightly increasing the mean service time of a grade. When an interior solution (of service rates) is optimal, the marginal values/costs of adjusting service rates must balance across two grades.

Lemma 1 presents necessary conditions of the optimal service rates for two-grade critically loaded systems. Building on these conditions, we can further show that if $-V'(\mu)\cdot\mu^2$ is strictly monotone, then none of these conditions will be satisfied. Thus, with strictly monotone $-V'(\mu)\cdot\mu^2$, the optimal service rates for a two-grade critically loaded system must either be on the boundary, namely, $(\bar{\mu},\underline{\mu})$, or degenerate to a single-grade critically loaded solution.

Proposition 4. *Consider the two-grade problem* (15) *with* $(w_1,w_2)=(0,0)$.

1. *If* $-V'(\mu)\cdot\mu^2$ *is strictly decreasing, then the optimal* $\boldsymbol{\mu}^*=(\bar{\mu},\underline{\mu})$ *and* $\mathcal{M}_2^{\mathrm{CL}}=\mathcal{R}^{\mathrm{CL}}(\bar{\mu},\underline{\mu})$.

2. *If* $-V'(\mu)\cdot\mu^2$ *is strictly increasing, then* $\mathcal{M}_2^{\mathrm{CL}}=\mathcal{M}_1^{\mathrm{CL}}$*; equivalently, the optimal two-grade critically loaded solution degenerates to the single-grade critically loaded solution.*

Proposition 4 shows that the shape of the reward function can determine the differentiation structure in a critically loaded system. With $\mu = 1/s$ and $U'(s) = -V'(\mu)\mu^2$, the monotonicity of $-V'(\mu)\mu^2$ in μ determines the curvature of $U(s)$ in s: if $-V'(\mu)\mu^2$ is decreasing, then U is convex in service time and it is highly valuable to spread customers across two different grades; if $-V'(\mu)\mu^2$ is increasing, then U is concave in service time and the optimal solution reduces to a single grade.

Managerially, convexity of rewards in service time is plausible in quality-based settings such as primary care, legal advisory services, and diagnostic services, where a short interaction with customers delivers limited values whereas a longer interaction generates substantially higher values. The above result suggests that in settings where the marginal value of longer service times rises sufficiently fast, the system manager should consider creating two grades with different service times, a fast standard grade and a slow premium grade.

4 Patience Time Distributions with Monotone Hazard Rates

This section builds on the analysis in §3.2 and derives sharper policy prescriptions for patience time distributions with monotone hazard rates. Specifically, we consider strictly increasing hazard rates (e.g., Erlang), constant hazard rates (i.e., exponential), and strictly decreasing hazard rates (e.g., hyper-exponential).[3]

The analysis below shows that the "at most two" structure in Proposition 1 is robust across these hazard-rate classes, but the exact forms of the service rate and offered wait pairs are sensitive to the hazard-rate structures. Specifically, with increasing or constant hazard rates, whenever two grades are optimal, the system must be critically loaded. In contrast, with decreasing hazard rates, a two-grade system can be optimal in an overloaded regime, applying a common offered wait to both grades.

4.1 Single-Grade Analysis

We start by analyzing the single-grade problem (11), where the total welfare of this grade is

$$\begin{aligned}\lambda_l r(\mu, w_l) + \lambda_h r(\mu, w_h) &= \sum_{j=l,h} \left[\min\{n_j\mu, \lambda_j\} V(\mu) - \lambda_j \int_0^{w_j} \bar{F}(y)dy\right] \\ &= V(\mu) \sum_{j=l,h} \min\{n_j\mu, \lambda_j\} - \sum_{j=l,h} \lambda_j \int_0^{w_j} \bar{F}(y)dy \\ &= V(\mu)\Lambda - V(\mu) \sum_{j=l,h} (\lambda_j - n_j\mu)^+ - \sum_{j=l,h} \lambda_j \int_0^{w_j} \bar{F}(y)dy. \end{aligned} \tag{16}$$

In (16), $\sum_{j=l,h}(\lambda_j - n_j\mu)^+$ and $\sum_{j=l,h} \lambda_j \int_0^{w_j} \bar{F}(y)dy$ are the steady-state abandonment rate and queue length of this grade, respectively. Thus, holding μ fixed, the scheduling policy will minimize a combined abandonment and queue-length cost. In fact, any non-idling policy will minimize the

[3] Monotone patience-time hazard rates are empirically relevant. For example, Brown et al. (2005), Li et al. (2018) document broadly decreasing patience-time hazard rates in call-center data.

abandonment cost and the optimal policy that minimizes the queue-length cost will depend on the hazard rate of the patience time distribution (Bassamboo and Randhawa 2016). This motivates us to consider patience time distributions with monotone hazard rates.

4.1.1 Increasing Patience Time Hazard Rates.

We first consider patience time distributions with increasing hazard rates. In this case, the optimal policy that minimizes the queue-length cost is to process customers under the Last Come First Served (LCFS) policy (Bassamboo and Randhawa 2016, Corollary 1). This policy is obtained by creating two subgrades with offered waits $(w_l, w_h) = (0, \infty)$ in the fluid solution. In other words, one subgrade is processed immediately upon arrival, and the other subgrade is never processed and abandons the system after waiting out their patience. To understand this policy, note that with increasing hazard rates, customers are more likely to abandon as they wait in queue. Thus, it is a good idea to serve them when they don't wait at all, or to never serve them and allow them to abandon after waiting out their patience.

Since any non-idling policy minimizes the abandonment cost, it is immediate that LCFS will minimize the combined abandonment and queue-length costs. With offered waits $(0, \infty)$, we can rewrite (16) as

$$\lambda_l r(\mu, w_l) + \lambda_h r(\mu, w_h) = NV(\mu) \cdot \mu - (\Lambda - N\mu)/\gamma.$$

Hence, the optimal single-grade welfare is given by

$$\mathcal{M}_1^* = \mathcal{M}_1^{\text{LCFS}} := \max_{\underline{\mu} \le \mu \le \mu_0} NV(\mu) \cdot \mu - (\Lambda - N\mu)/\gamma. \tag{17}$$

4.1.2 Decreasing Patience Time Hazard Rates.

Next, consider patience time distributions with decreasing hazard rates. In this case, the optimal policy that minimizes the queue-length cost is FCFS (Bassamboo and Randhawa 2016, Corollary 1). Since any non-idling policy can minimize the abandonment cost, it is immediate that FCFS will minimize the combined abandonment and queue-length costs. Under FCFS, we can rewrite (16) as

$$\lambda_l r(\mu, w_l) + \lambda_h r(\mu, w_h) = NV(\mu) \cdot \mu - \Lambda \int_0^{\bar{F}^{-1}\left(\frac{N\mu}{\Lambda}\right)} \bar{F}(y) dy.$$

Hence, the optimal single-grade welfare is given by

$$\mathcal{M}_1^* = \mathcal{M}_1^{\text{FCFS}} := \max_{\underline{\mu} \le \mu \le \mu_0} NV(\mu) \cdot \mu - \Lambda \int_0^{\bar{F}^{-1}\left(\frac{N\mu}{\Lambda}\right)} \bar{F}(y) dy. \tag{18}$$

Note that unlike (17), the optimal service rate in (18) depends on the entire patience distribution.

4.1.3 Exponential Patience Time: Constant Hazard Rates.

Under exponential patience times, we can leverage the memoryless property to convert any queue-length-based objective to an equivalent abandonment-based objective, so any non-idling policy (including FCFS and LCFS) will optimize (16). In this case, we have $\mathcal{M}_1^* = \mathcal{M}_1^{\text{LCFS}} = \mathcal{M}_1^{\text{FCFS}}$.

4.2 Two-Grade Analysis

We next analyze the two-grade optimization problem (13). Recall that in principle we should solve (13) in two stages by optimizing the service rates and offered waits in a sequential manner. Proposition 2 provides optimality conditions of the (second-stage) offered waits under given service rates. However, there are in general too many candidate offered waits to consider and this creates challenges in exhausting all those offered waits to establish optimality. We thus follow the preceding analysis and focus on patience time distributions with monotone hazard rates. This allows us to considerably refine candidate offered waits.

Lemma 2. *Suppose the hazard rate of the patience time distribution is strictly monotone. Then, under any service rates* (μ_1, μ_2) *such that* $\mu_1 > \mu_2$*, the optimal* (w_1, w_2) *cannot be such that* $w_1 \neq w_2$ *and* $0 < w_1, w_2 \leq \infty$*.*

The condition in Lemma 2 requires the hazard rate of the patience distribution to be *strictly* monotone (e.g., Erlang and hyper-exponential), so it will not apply to exponential patience distributions. (We will consider exponential patience distributions in §4.2.3.) Using Lemma 2, for patience distributions with monotone hazard rates, we only need to consider offered waits (w_1, w_2) in the form of (w, w), $(0, w_2)$, $(w_1, 0)$, $(0, 0)$, $(0, \infty)$, and $(\infty, 0)$ for $0 < w, w_1, w_2 < \infty$. Among them, if $(w_1, w_2) = (0, \infty)$, we have $\boldsymbol{\lambda} = (N\mu_1, \Lambda - N\mu_1)$ (this solution is feasible if and only if $\mu_1 \leq \mu_0$). In this case, grade 1 is served completely and grade 2 is not served at all. Similarly, if $(w_1, w_2) = (\infty, 0)$, we have $\boldsymbol{\lambda} = (\Lambda - N\mu_2, N\mu_2)$. In this case, grade 2 is served completely and grade 1 is not served at all. In both cases, only one grade is effectively served. This is equivalent to creating a single grade and processing that grade under LCFS. So, the optimal welfare under $(w_1, w_2) = (0, \infty)$ and $(w_1, w_2) = (\infty, 0)$ is the same as $\mathcal{M}_1^{\text{LCFS}}$. Hence, to obtain a non-degenerate two-grade policy, we can further restrict analysis to (w, w), $(0, w_2)$, $(w_1, 0)$, and $(0, 0)$. In other words, the optimal two-grade policy should *either strictly prioritize one grade or set the same offered wait for both grades.*

4.2.1 Increasing Patience Time Hazard Rates.

We first consider patience-time distributions with increasing hazard rates. For this class, the next result shows that the candidate offered wait vector reduces to $(0, 0)$.

Lemma 3. *Suppose the patience time distribution has a strictly increasing hazard rate. Then, a solution with a finite and positive* w_1 *or* w_2 *cannot be optimal.*

An intuitive explanation for Lemma 3 is that with increasing hazard rates, it is never optimal to process a grade using a finite and positive offered wait, because doing so is always dominated by LCFS. Thus, the optimal offered waits for a non-degenerate two-grade policy must be (0,0). This allows us to focus on critically loaded systems to establish the optimal service rate(s). It also presents a case in which *customers are not differentiated on their offered waits (because both are zero) but on their service rates only.* The next result follows immediately from Lemma 3.

Proposition 5. *Suppose the patience time distribution has a strictly increasing hazard rate. Then, the optimal system welfare is* $\max\{\mathcal{M}_1^{\mathrm{LCFS}}, \mathcal{M}_2^{\mathrm{CL}}\}$*, where* $\mathcal{M}_1^{\mathrm{LCFS}}$ *is defined in* (17) *and* $\mathcal{M}_2^{\mathrm{CL}}$ *is defined in* (15)*. Moreover, any nondegenerate two-grade optimal solution must be critically loaded, with offered waits* $(w_1, w_2) = (0, 0)$*.*

Proposition 5 shows that, under increasing hazard rates, the optimal policy either creates one grade and processes it under LCFS, possibly in an overloaded regime, or creates two grades and processes each grade under FCFS in a critically loaded regime. Thus, whenever a nondegenerate two-grade system is optimal, it is critically loaded. The choice between these two depends on the comparison between $\mathcal{M}_1^{\mathrm{LCFS}}$ and $\mathcal{M}_2^{\mathrm{CL}}$, which is governed by the reward function V and the mean patience time $1/\gamma$. The numerical study in §4.3 provides examples in which either structure can dominate.

The next result characterizes the optimal policy when $\mu\,[V(\mu) + 1/\gamma]$ is strictly monotone.

Corollary 2. *Suppose the patience time distribution has a strictly increasing hazard rate.*

1. *If* $\mu\,[V(\mu) + 1/\gamma]$ *is strictly increasing, then the optimal welfare is the larger one between* $\mathcal{M}_1(\mu_0)$ *and* $\mathcal{M}_2^{\mathrm{CL}}$ *in* (15)*.*
2. *If* $\mu\,[V(\mu) + 1/\gamma]$ *is strictly decreasing, then the optimal welfare is* $\mathcal{M}_1(\underline{\mu})$ *and the optimal policy should have only one grade with service rate* $\underline{\mu}$ *and two offered waits* $(0, \infty)$*.*

To establish Corollary 2, we show in the proof that a necessary condition for $\mathcal{M}_2^{\mathrm{CL}}$ to dominate is that the optimal service rates (μ_1, μ_2), with $\mu_1 > \mu_2$, satisfy $\mu_1[V(\mu_1) + 1/\gamma] \geq \mu_2[V(\mu_2) + 1/\gamma]$. Equivalently, the faster grade must be prioritized under the optimal scheduling rule (Bassamboo et al. 2023, Proposition 4). This condition is automatically satisfied when $\mu[V(\mu)+1/\gamma]$ is increasing and is violated when $\mu[V(\mu) + 1/\gamma]$ is strictly decreasing.

4.2.2 Decreasing Patience Time Hazard Rate.

We next consider patience time distributions with decreasing hazard rates. The candidate offered waits can be refined as follows.

Lemma 4. *Suppose the patience distribution has a strictly decreasing hazard rate. Then, for any* (μ_1, μ_2) *such that* $\mu_1 > \mu_2$*, the optimal* (w_1, w_2) *cannot be of the form* $(w_1, 0)$ *or* $(0, w_2)$ *for* $0 < w_1, w_2 < \infty$*.*

Using Lemma 4, for patience distributions with decreasing hazard rates, we only need to consider offered waits in the form of $(0, 0)$ and (w, w) for $0 < w < \infty$. (Note that unlike Lemma 3, Lemma 4 does not rule out the latter.) Interestingly, in both cases, customers are differentiated on their service rates but not on their waiting times. This presents another example in which *waiting time differentiation is moot but service rate differentiation takes an exclusive effect.*

Combining the two cases in a unified formulation with a common offered wait $w \geq 0$, define[4]

$$\mathcal{M}_2^{\mathrm{OL}} := \sup_{\substack{\underline{\mu} \leq \mu_2 \leq \mu_1 \leq \bar{\mu},\ w \geq 0 \\ \lambda_1, \lambda_2 \geq 0}} \bar{F}(w)\left[\lambda_1 V(\mu_1) + \lambda_2 V(\mu_2)\right] - \Lambda \int_0^w \bar{F}(y)\, dy \tag{19}$$

$$\text{s.t.} \quad \lambda_1 + \lambda_2 = \Lambda, \quad \bar{F}(w)\left(\lambda_1/\mu_1 + \lambda_2/\mu_2\right) = N.$$

It is evident that $\mathcal{M}_2^{\mathrm{OL}} \geq \mathcal{M}_2^{\mathrm{CL}}$ since $\mathcal{M}_2^{\mathrm{CL}}$ is a special case of $\mathcal{M}_2^{\mathrm{OL}}$ with $w = 0$. Then, the equality is attained only when the optimal $\mathcal{M}_2^{\mathrm{OL}}$ has $w^* = 0$. The next result follows from Lemma 4.

Proposition 6. *Suppose the patience time distribution has a strictly decreasing hazard rate. Then, the optimal system welfare is $\mathcal{M}_2^{\mathrm{OL}}$ in* (19)*. The optimizer may be critically loaded, overloaded, or degenerate to the single-grade FCFS solution.*

Proposition 6 shows that decreasing hazard rates can support *overloaded* two-grade systems, in contrast to the increasing-hazard case in Proposition 5. In such systems, both grades share the same offered wait, so waiting time differentiation is absent across grades; the differentiation occurs through service rates. The next result identifies when the common offered wait is strictly positive.

Lemma 5. *Let (μ_1^*, μ_2^*) with $\mu_1^* > \mu_2^*$ be the optimal service rates in* (19)*. The corresponding optimal offered wait satisfies $w^* > 0$ if and only if $w = w^*$ satisfies*

$$-\frac{V(\mu_1^*)\mu_1^* - V(\mu_2^*)\mu_2^*}{\mu_1^* - \mu_2^*} = \frac{1}{H(w)} \quad \text{and} \quad \bar{F}^{-1}\left(\frac{N\mu_1^*}{\Lambda}\right) \leq w \leq \bar{F}^{-1}\left(\frac{N\mu_2^*}{\Lambda}\right), \tag{20}$$

where $H(\cdot)$ is the hazard rate and $\bar{F}^{-1}(x) \equiv 0$ for $x > 1$.

We leverage Lemma 5 to provide a complete characterization of the optimal two-grade systems provided that $-V'(\mu) \cdot \mu^2$ is strictly decreasing.

Corollary 3. *Suppose the patience distribution has a strictly decreasing hazard rate and $-V'(\mu) \cdot \mu^2$ is strictly decreasing. Then, the optimal $\mathcal{M}_2^{\mathrm{OL}}$ in* (19) *is achieved under $\boldsymbol{\mu} = (\bar{\mu}, \underline{\mu})$ and the optimal $w^* > 0$ if and only if $-\frac{V(\bar{\mu})\cdot\bar{\mu} - V(\underline{\mu})\cdot\underline{\mu}}{\bar{\mu} - \underline{\mu}} \cdot H(0) > 1$ and $-\frac{V(\bar{\mu})\cdot\bar{\mu} - V(\underline{\mu})\cdot\underline{\mu}}{\bar{\mu} - \underline{\mu}} \cdot H\left(\bar{F}^{-1}\left(\frac{N\underline{\mu}}{\Lambda}\right)\right) < 1$.*

4.2.3 Exponential Patience Times: Constant Hazard Rate.

Finally, we consider exponential patience times with constant hazard rates. Despite the commonly conceived simplicity of exponential distributions due to their memoryless property, such property in fact precludes offered wait refinements that we adopt for patience distributions with monotone hazard rates. (Recall that Lemma 2 requires a strictly monotone hazard rate and therefore does not apply to the constant-hazard exponential case.) Thus, we must work with the general formulation (13) to identify the optimal solution.

Indeed, the cases of (w_1, ∞) and (∞, w_2) for $0 < w_1, w_2 < \infty$ which are ruled out by Lemma 2 for patience distributions with monotone hazard rates, can be valid offered waits for exponential

[4]The superscript 'OL' stands for 'overloaded.'

patience distributions. They, along with other candidate offered waits (w_1, w_2) in the form of (w, w), $(0, w_2)$, $(w_1, 0)$, $(0, 0)$, $(0, \infty)$, and $(\infty, 0)$ for $0 < w, w_1, w_2 < \infty$, constitute the feasible set of candidate offered waits. To circumvent the difficulty of considering the entire set, we alternatively optimize over arrival rates $\boldsymbol{\lambda}$ allocated to each grade. This allows us to establish the following result, which surprisingly parallels Proposition 5 for strictly increasing hazard rates.

Proposition 7. *Suppose the patience time is exponentially distributed. Then,* $\mathcal{M}_1^* = \mathcal{M}_1^{\mathrm{LCFS}} = \mathcal{M}_1^{\mathrm{FCFS}}$. *Further, the optimal welfare is the larger one among* $\mathcal{M}_1^*$ *and* $\mathcal{M}_2^{\mathrm{CL}}$ *in* (15).

The parallel arises because, note that the optimal (second-stage) scheduling policies for exponential patience distributions and patience distributions with increasing hazard rates are very similar (Bassamboo et al. 2023). Specifically, the marginal benefit of allocating an incremental capacity to grade i is given by $\mu_i[V(\mu_i) + 1/\gamma]$ under these patience distributions and the optimal scheduling policy prioritizes grades in descending order of this index.

The next result follows in a similar fashion to Corollary 2.

Corollary 4. *Suppose the patience time is exponentially distributed.*

1. *If* $\mu\,[V(\mu) + 1/\gamma]$ *is strictly increasing, then the optimal welfare is the larger one between* $\mathcal{M}_1(\mu_0)$ *and* $\mathcal{M}_2^{\mathrm{CL}}$ *in* (15).
2. *If* $\mu\,[V(\mu) + 1/\gamma]$ *is strictly decreasing, then the optimal welfare is* $\mathcal{M}_1(\underline{\mu})$ *and the optimal policy should have only one grade* $\underline{\mu}$.

4.2.4 Summary of Optimal Policies

For patience time distributions with monotone hazard rates, we summarize in Table 2 below their corresponding optimal policies. In general, the hazard rate of patience time determines not only the optimal within-grade scheduling rule (Bassamboo and Randhawa 2016), but also whether service rate differentiation should exist, and if yes, in what regimes. With strictly increasing or constant hazard rates of patience times, any two-grade system must be critically loaded; overloaded systems, if optimal, can have only one grade. With strictly decreasing hazard rates of patience times, two-grade differentiation can exist in an overloaded regime, and both grades share a common offered wait.

Table 2: Optimal policies for patience time distributions with different hazard rate structures

Hazard rate of patience	Single-grade policy	Two-grade policy (non-degenerate)	Overload & Two-grade
Strictly increasing	LCFS	FCFS within each grade, $(w_1, w_2) = (0, 0)$	Not possible
Constant	Any non-idling	FCFS within each grade, $(w_1, w_2) = (0, 0)$	Not possible
Strictly decreasing	FCFS	Common offered wait, $(w_1, w_2) = (w, w)$	Possible

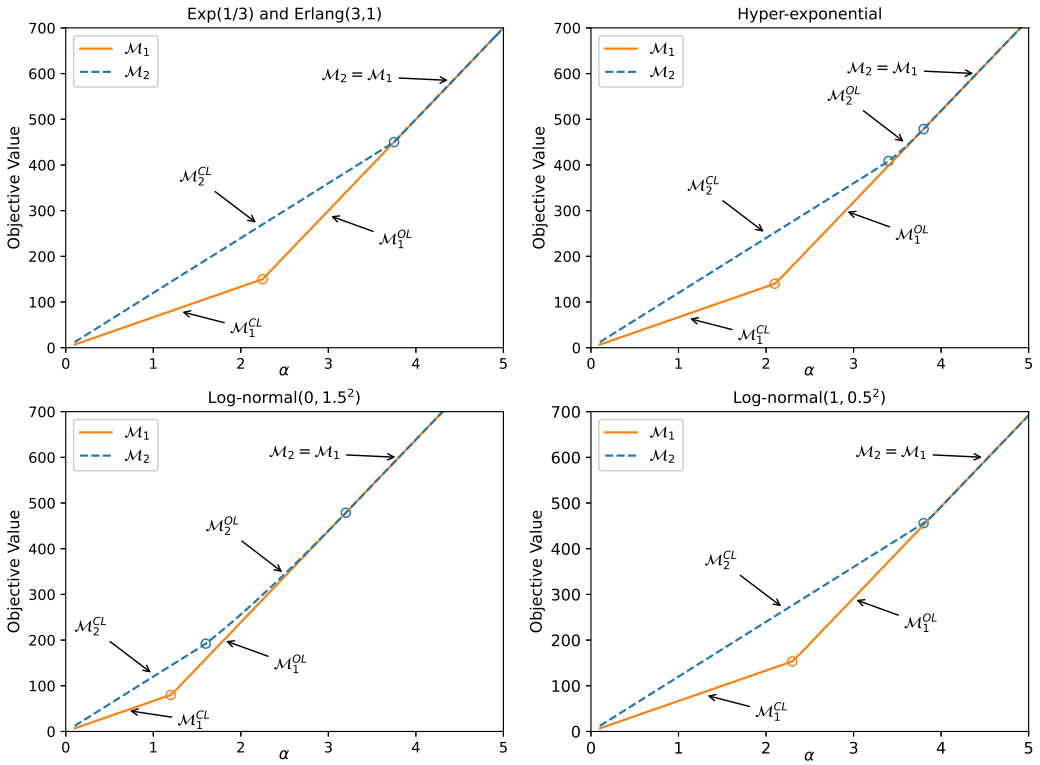
4.3 Numerical Example

We provide numerical examples to demonstrate the benefits of service rate differentiation as well as regime changes under different model parameters. We focus on the fluid optimization problem to

present our numerical results (results in finite stochastic systems can be found in §5). To ensure nontrivial service rate differentiation (see Table 1), we use the reward function $V(\mu) = \alpha/\mu^2$ and vary α. For this specification, $V(\mu)\mu = \alpha/\mu$ is decreasing and convex, so Proposition 3 does not rule out a nondegenerate two-grade optimum. We set $(\Lambda, N, c) = (150, 100, 1)$ and $(\underline{\mu}, \bar{\mu}) = (0.5, 2.5)$. We consider five patience time distributions: exponential with rate $1/3$, Erlang with shape 3 and rate 1, hyper-exponential with rates $1/2$ and $1/4$ mixed with equal probabilities, and two log-normal distributions whose log-patience times are distributed as $N(0, 1.5^2)$ and $N(1, 0.5^2)$. These distributions have very similar means: 3.08 for the two log-normal distributions and 3 for the other three. Their hazard-rate shapes, however, differ substantially, leading to different policy prescriptions.

For each patience time distribution, we solve (11) and (13). Figure 1 plots the optimal single-grade and two-grade values, $\mathcal{M}_1^*$ and $\mathcal{M}_2^*$, as functions of α; hollow circles mark regime transitions. As α increases, the reward from slower, higher-quality service becomes more important relative to waiting cost. Since $V(\mu)\mu = \alpha/\mu$ is decreasing in μ, the manager has a stronger incentive to choose lower service rates, which, as α increases, can move the system from critically loaded to overloaded.

Figure 1: Optimal values of $\mathcal{M}_1^*$ and $\mathcal{M}_2^*$ under different patience time distributions.



The benefit of service rate differentiation can be measured by the relative gap $(\mathcal{M}_2^* - \mathcal{M}_1^*)/\mathcal{M}_1^*$, which is especially large for small values of α. It reaches roughly 80% in the critically loaded regime, where all fluid customers enter service without delay and the queue length is zero, regardless of the patience time distribution; hence the welfare gain comes entirely from service rate differentiation.

4.3.1 Discussion

Exponential and Erlang patience distributions yield the same optimal grade allocation and scheduling policy. In the critically loaded regime, the optimal single-grade welfare is $\mathcal{M}_1^{\mathrm{CL}} = \mathcal{M}_1(\mu_0)$, while the optimal two-grade welfare is $\mathcal{M}_2^{\mathrm{CL}}$. As α increases, the optimal solution eventually enters an overloaded regime. In this case, the two-grade policy degenerates to a single-grade solution, leading to $\mathcal{M}_2^* = \mathcal{M}_1^* = \mathcal{M}_1^{\mathrm{OL}}$. In other words, two-grade overloaded systems are never optimal, consistent with our prescriptions in Propositions 5 and 7. In contrast, for hyper-exponential patience times, the optimal solution exhibits a distinct intermediate region in which a two-grade policy in the overloaded regime is optimal. These results, collectively, suggest how the entire patience time distribution, beyond its mean, affects the optimal policy and system regimes.

The two log-normal patience distributions have non-monotone hazard rates, so the results in §4.2 developed for monotone hazard rates do not directly apply. We thus solve (11) and (13) numerically. The obtained policies, however, resemble those in the monotone-hazard-rate cases. The log-normal distribution with parameters $(0, 1.5^2)$ behaves similarly to the hyper-exponential case: an overloaded two-grade policy can be optimal. In contrast, the log-normal distribution with parameters $(1, 0.5^2)$ resembles the exponential and Erlang cases: an overloaded two-grade policy never appears optimal.

This distinction can be understood through the queue-length-minimizing policy for a single-grade *overloaded* system under log-normal patience times. The optimal policy is either FCFS or represented by a two-offered-wait policy $(0, \widehat{w})$, where $\widehat{w}$ solves Bassamboo and Randhawa (2016, Equation 24). FCFS is optimal when the FCFS offered wait $\bar{w}$ exceeds $\widehat{w}$; otherwise, the grade is split into two subgrades. However, Proposition 1 and Corollary 1 suggest that, when the optimal policy uses two service grades, no grade needs to be further split. With log-normal patience times, this implies that either the offered wait of a grade is zero (not overloaded) or the offered wait of this grade processed under FCFS is greater than the corresponding threshold $\widehat{w}$. For log-normal$(0, 1.5^2)$, the threshold is $\widehat{w} = 0.35$, so the condition $\bar{w} > \widehat{w}$ can hold in the parameter region we consider. This explains why the optimal policy resembles the hyper-exponential case. For log-normal$(1, 0.5^2)$, the threshold is $\widehat{w} = 31.8$, making $\bar{w} > \widehat{w}$ highly unlikely in our parameter region. Accordingly, the optimal solution does not exhibit an overloaded two-grade phase and thus resembles the exponential and Erlang cases.

5 Implementing the Fluid Solution in Stochastic Queueing Systems

The fluid analysis in the preceding sections provides a guideline for designing implementable policies in stochastic queueing systems. Guided by the fluid formulation, we propose an $\mathcal{S}$ policy that implements service rate differentiation at arrival: each customer is assigned to a service grade upon arrival, and scheduling decisions then use this grade information together with customers' elapsed waiting times. An equivalent waiting-time-first implementation is discussed in Appendix A.

5.1 The $\mathcal{S}$ Policy

The $\mathcal{S}$ policy implements the fluid formulation in a stochastic queueing system. Upon arrival, each customer is assigned to a service grade, and each grade is associated with a service rate from the fluid solution. Operationally, assigning a customer to grade i means that, conditional on entering service, her service time is drawn from the distribution $G_{\mu_i^*}$ associated with the chosen service rate μ_i^*. Scheduling then uses the assigned grade and the customer's elapsed waiting time, as follows:

1. (Single-grade policy) If the fluid solution has only one grade μ^*, then set the service rate equal to μ^* for all arriving customers.

   (a) If the fluid solution has a single offered wait w^*, then process all customers under FCFS.

   (b) If the fluid solution has two different offered waits $w_l^* < w_h^*$, then process customers using the following time-in-queue (TIQ) policy (Bassamboo and Randhawa 2016):

      (i) First, process customers whose elapsed waiting times exceed w_h^*, in FCFS order.

      (ii) Next, process customers whose elapsed waiting times are below w_l^*, in FCFS order.

      (iii) Finally, process the remaining customers under LCFS.

2. (Two-grade policy) If the fluid solution has two grades $\mu_1^* \neq \mu_2^*$ with corresponding arrival rates $(\lambda_1^*, \lambda_2^*)$ and offered waits (w_1^*, w_2^*). Then,

   (a) (Grade allocation) Split incoming customers into two grades with arrival rates $(\lambda_1^*, \lambda_2^*)$, that is, assign each customer to grade 1 with probability λ_1^*/Λ and to grade 2 otherwise. Set the service rates of the two grades equal to (μ_1^*, μ_2^*).

   (b) (Scheduling) Process customers using the mostly-FCFS policy (Bassamboo et al. 2023).[5]

      (i) If $w_1^* = w_2^* = 0$, then use a static priority rule. At each service completion, give priority to the grade with the larger marginal capacity value β_i; within each grade, serve customers in FCFS order. Under the separable welfare function (3), the marginal value β_i can be calculated as follows

$$\beta_i = \begin{cases} V(\mu_i^*)\mu_i^* + \frac{\mu_i^*}{H(w_{i,l})} & \text{if } w_{i,l} > 0, \\ V(\mu_i^*)\mu_i^* + \frac{\mu_i^*}{H(w_{i,h})} & \text{if } w_{i,l} = 0, w_{i,h} < \infty, \text{ for } i = 1, 2, \\ V(\mu_i^*)\mu_i^* + \frac{\mu_i^*}{\gamma} & \text{if } w_{i,l} = 0, w_{i,h} = \infty, \end{cases}$$

      where H is the hazard rate function of the patience distribution, $(w_{i,l}, w_{i,h})$ are the limiting offered waits[6] to optimally process grade i under service rate μ_i^*, arrival rate λ_i^*, and capacity $n \uparrow \frac{\lambda_i^*}{\mu_i^*}$.

      Prioritize the grade with a higher β and process that grade under FCFS. Then, process the remaining grade under FCFS.

[5] The mostly-FCFS policy reduces to an all-FCFS policy in our two-grade problem because customers within each grade are processed under FCFS.

[6] These limits exist and are well defined; see Bassamboo et al. 2023, Lemma 2

(ii) If exactly one offered wait is zero, say $w_i^* = 0$ and $w_{-i}^* > 0$, then give priority to grade i; within each grade, serve customers in FCFS order.

(iii) If both w_1^* and w_2^* are positive and finite, then give priority to customers whose elapsed waiting time exceeds the fluid offered wait target of their grade. Among these customers, serve the one with the longest elapsed waiting time. If no customer exceeds her grade-specific target, serve the remaining customers in FCFS order.

Note that service rate differentiation is not relevant in single-grade systems, in which case, all customers are served at the same service rate under either FCFS or TIQ, both ensuring that those who get served meet the target offered waits.

In contrast, service rate differentiation is central in two-grade systems. Assigning different service rate grades to ex ante homogeneous customers creates the strategic heterogeneity needed for scheduling, and the offered waits obtained from (13) determine how idle servers should be routed across grades. Because customers within each grade are processed under FCFS, a key scheduling decision is which grade should be prioritized over the other. The rules in Step 2(b)(i)–(iii) address this decision in three offered wait configurations, utilizing the following notion of *the marginal value of capacity*:

- Step 2(b)(ii) applies when exactly one grade has zero offered wait. In the fluid solution, the zero-wait grade is fully served, whereas the positive-wait grade is only partially served; hence the zero-wait grade has priority.

- Step 2(b)(iii) applies when both grades have positive offered waits. In this case, Proposition 2 Part 2 suggests equal marginal values of capacity across two active grades, so we propose implementing priority based on customers' elapsed waiting time relative to their grade-specific offered wait target: customers who have waited more than their targets are served first, and priority is given to such a customer with longest waiting.

- Finally, Step 2(b)(i) applies when both offered waits are zero. At the fluid scale, the system is critically loaded, and any non-idling rule can achieve zero offered waits for both grades. In the stochastic system, however, finite-system fluctuations can make the marginal value of an idle server differ across two grades. We thus define an index β to guide the tie-breaking rule: prioritize the grade with the higher marginal value of capacity β, and serve customers within that grade under FCFS.

To implement the $\mathcal{S}$ policy, the system should keep track of two types of information: the grade label assigned to a customer at arrival and that customer's elapsed waiting time in queue. The grade information determines the customer's service time distribution when she enters service, while the elapsed waiting time determines whether the customer has reached the fluid offered wait threshold corresponding to her grade. In this sense, the $\mathcal{S}$ policy is implemented as a randomized grade-assignment rule at arrival jointly with a priority rule based on grade labels and elapsed waiting times. Alternatively, Appendix A proposes an equivalent waiting-time-first implementation

at the fluid scale. Under that implementation, customers are not segregated upon arrival; instead, they are first differentiated by their elapsed waiting time, and when they enter service, each of them is assigned a (possibly different) service rate. We show that these two implementations are fluid-equivalent, but their finite-system performance need not be identical; in fact, our numerical results suggest a slight performance advantage of the $\mathcal{S}$ policy in stochastic systems and we thus focus on illustrating this policy in this section.

5.2 Performance of Proposed Policy for Markovian Systems

For Markovian systems, the patience time distributions are exponential. Following the preceding discussion, if the fluid solution has only one grade, then all customers are processed under FCFS. If the fluid solution has two grades, then these grades under the $\mathcal{S}$ policy are prioritized in descending order of $V(\mu_i) \cdot \mu_i + \mu_i/\gamma$ and customers of each grade are processed under FCFS. We next provide a theoretical result on the performance of $\mathcal{S}$ policy for Markovian systems. To formally state the result, we introduce the following notation. We denote the total arrival rate by Λ and fix $\mu_0 = \Lambda/N$ by setting the number of servers $N = \Lambda/\mu_0$. We denote the system welfare using the optimal policy (among all non-anticipating policies) by K^*_Λ, the system welfare under the $\mathcal{S}$ policy by K_Λ, and the optimal objective value of the fluid optimization problems (11) and (13) by $\mathcal{M}^*(N, \Lambda)$.

Proposition 8. *Assume that the welfare function is separable. If the inter-arrival, service, and patience times are all exponentially distributed, then the $\mathcal{S}$ policy has $\mathcal{O}(\sqrt{\Lambda})$ performance. That is, there exists finite $A > 0$ such that*

$$K^*_\Lambda - K_\Lambda \leq A\sqrt{\Lambda} \quad \text{as } \Lambda \to \infty.$$

Moreover, if optimal value $\mathcal{M}^*(N, \Lambda)$ is achieved under (11) with $w^* > 0$, then our proposed policy has $o(1)$ performance. That is,

$$\left|K^*_\Lambda - K_\Lambda\right| \to 0 \quad \text{as} \quad \Lambda \to \infty.$$

Further, the optimal system welfare is upper bounded by the optimal objective value of the fluid optimization problem (11) and (13). That is, $K^*_\Lambda \leq \mathcal{M}^*(N, \Lambda)$.

The above result shows that for Markovian systems, the gap between the $\mathcal{S}$ policy and the optimal policy is bounded by the square root of the system size times a constant. This can be understood by noting that the order of stochastic fluctuation in a critically loaded system is within the square root of the system size. However, when an overloaded system is optimal ($w^* > 0$ in a single-grade system), the performance gap between our policy and the optimal policy is not only bounded but tends to zero as the system size grows.

5.3 Numerical Examples

We next use simulations to illustrate the performance of our proposed policies in stochastic systems. We follow the parameters in §4.3 and simulate a system with Poisson arrivals at rate $\Lambda = 150$. The

system has $N = 100$ agents, each processing work at a unit rate. The feasible set of service rates for each grade is $[0.5, 2.5]$, and the service times of a grade with service rate μ are exponentially distributed with mean $1/\mu$. The reward function $V(\mu) = \alpha/\mu^2$, and we vary $\alpha \in [2, 4]$ to obtain different policies. The patience distributions we consider are exponential distribution with rate $1/3$, Erlang distribution with shape 3 and rate 1, hyper-exponential distribution with rates $1/2$ and $1/4$ and equal probability of occurrence of each rate, and two log-normal distributions, log-normal$(0, 1.5^2)$ and log-normal$(1, 0.5^2)$.

We report the average system welfare (per time unit) by simulating each queueing system for 10,000 time units and taking an average of 20 independent runs; the 95% confidence interval half-width is less than 0.2% of the reported values for all cases we consider.

5.3.1 Benefits of Service Rate Differentiation.

We compare the performance of systems with and without service rate differentiation using simulations. Both systems utilize optimal scheduling: the single-grade system follows FCFS or TIQ under the optimal (single-grade) service rate, and the two-grade system follows the $\mathcal{S}$ policy for joint service rate and waiting time differentiation. Table 3 reports the results under various patience time distributions and shows that gains from service rate differentiation can be as much as 90%, suggesting significant improvement from this dimension.

Table 3: Benefits of Service Rate Differentiation: Simulation of $\mathcal{S}$ policy, $V(\mu) = \alpha/\mu^2, \Lambda = 150, N = 100$, $\mu \in [0.5, 2.5]$.

	$\alpha = 2$		$\alpha = 3$	
Patience Distribution	Single-Grade	Two-Grade	Single-Grade	Two-Grade
Exp(1/3)	119.2	+85.5%	300.3	+11.6%
Erlang(3,1)	114.3	+90.9%	299.4	+11.8%
Hyper-exponential	119.2	+85.7%	318.8	+5.1%
Log-normal$(0, 1.5^2)$	238.1	+7.3%	438.4	+0.2%
Log-normal$(1, 0.5^2)$	113.8	+89.2%	291.6	+14.0%

5.3.2 Accuracy of Fluid Approximations.

We present in Table 4 our simulation results of the $\mathcal{S}$ policy and compare them with those obtained from the fluid solution. For $\alpha = 2$ and 3, the queue-length cost has a similar magnitude to service rewards. A two-grade critically loaded system is optimal for all patience distributions except for log-normal$(0, 1.5^2)$ (see Figure 1 for illustration). For most of these patience distributions, we observe similar gaps between our policy and the fluid solution. (The fluid solution is the same for these patience distributions because the fluid system is critically loaded and customers don't abandon.) However, for the patience distribution log-normal$(0, 1.5^2)$, the optimal policy operates a two-grade

overloaded system, so there is only a tiny gap between our policy and the fluid solution. For $\alpha = 4$, the queue-length cost is not comparable to service rewards. This leads to an optimal single-grade system in the overloaded regime for all patience distributions (see Figure 1 for illustration). So, the gap between our policy and the fluid solution is minimal for all patience distributions.

Table 4: Accuracy of Fluid Approximations: $V(\mu) = \alpha/\mu^2, \Lambda = 150, N = 100, \mu \in [0.5, 2.5]$.

	$\alpha = 2$		$\alpha = 3$		$\alpha = 4$	
Patience Distribution	Simulation	Fluid	Simulation	Fluid	Simulation	Fluid
Exp(1/3)	221.1	+8.5%	335	+7.5%	499.6	+0.1%
Erlang(3,1)	218.2	+10.0%	334.6	+7.6%	499.8	+0.0%
Hyper-exponential	221.3	+8.5%	335.1	+7.4%	517.9	+0.1%
Log-normal$(0, 1.5^2)$	255.4	+0.3%	439.3	+0.0%	638.5	+0.0%
Log-normal$(1, 0.5^2)$	215.3	+11.5%	332.5	+8.3%	491.8	+0.0%

6 Extension to Heterogeneous Customers

In this section, we extend our previous analysis to systems with heterogeneous customers. We show how the main structure of our result developed for homogeneous customers can qualitatively extend to heterogeneous customers with ex ante differentiated primitives such as arrival rates, patience distributions, and welfare functions.

Consider m customer classes. For class $i = 1, \ldots, m$, let Λ_i denote the arrival rate, F_i denote the patience time distribution, and $\bar{F}_i$ denote its survival function. Let $r_i(\mu, w)$ denote the welfare generated by a class-i customer who is processed at service rate μ with offered wait w. The system has N agents, each processing work at a unit rate.

We formulate the fluid optimization problem of heterogeneous customers in two stages. In the first stage, the manager chooses the number of service grades within each class, the service rate of each grade, and the allocation of class-i arrivals across these grades. Let $J(i)$ denote the number of service grades created for class i. Let $\boldsymbol{\mu}_i = (\mu_{i,1}, \ldots, \mu_{i,J(i)})$ and $\boldsymbol{p}_i = (p_{i,1}, \ldots, p_{i,J(i)})$, where $p_{i,j}$ is the fraction of class-i arrivals assigned to grade j. Define $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m)$ and $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m)$. The *first-stage grade-allocation problem* is

$$\max_{\boldsymbol{\mu}, \boldsymbol{p}} \ \mathcal{R}(\boldsymbol{\mu}, \boldsymbol{p}), \tag{21}$$

where $\mathcal{R}(\boldsymbol{\mu}, \boldsymbol{p})$ is the optimal welfare achieved by the *second-stage scheduling problem*.

To state the second-stage problem, let $K(i, j)$ denote the number of subgrades created within service grade j of class i, where each subgrade is processed under FCFS and is associated with a distinct offered wait. Let $\lambda_{i,j,k}$, $w_{i,j,k}$, and $n_{i,j,k}$ denote the arrival rate, offered wait, and capacity

allocated to subgrade k of grade j in class i. Then

$$\mathcal{R}(\boldsymbol{\mu}, \boldsymbol{p}) = \sup_{\{K(i,j),\, n_{i,j,k},\, w_{i,j,k},\, \lambda_{i,j,k}\}} \quad \sum_{i,j,k} \lambda_{i,j,k} r_i(\mu_{i,j}, w_{i,j,k}) \tag{22}$$

$$\text{s.t.} \quad \sum_{k=1}^{K(i,j)} \lambda_{i,j,k} = \Lambda_i p_{i,j}, \quad \forall (i,j),$$

$$\lambda_{i,j,k} \bar{F}_i(w_{i,j,k}) \le n_{i,j,k} \mu_{i,j}, \quad \forall (i,j,k),$$

$$\sum_{i,j,k} n_{i,j,k} \le N. \tag{23}$$

Proposition 9. *There exists an optimal solution to* (21) *and* (22) *where at most one customer class uses two service rate and offered wait pairs, and every other class uses one service rate and offered wait pair.*

Proposition 9 is an analogue of Proposition 1 in the heterogeneous-customer case. Interestingly, it shows that the benefit of service rate differentiation can be fully achieved by *splitting at most one customer class*. This adds important practical implementability to the differentiation policy. Proposition 9 further generalizes the structural property of the optimal policy for homogeneous customers in Proposition 1. Indeed, with heterogeneous customers, ex ante heterogeneity does not fully eliminate the parsimonious form of the optimal solution; rather, it extends the prescription from "at most two active pairs" under homogeneous customers to "at most one active pair deviating from single-grade FCFS" under the more general heterogeneous customers.

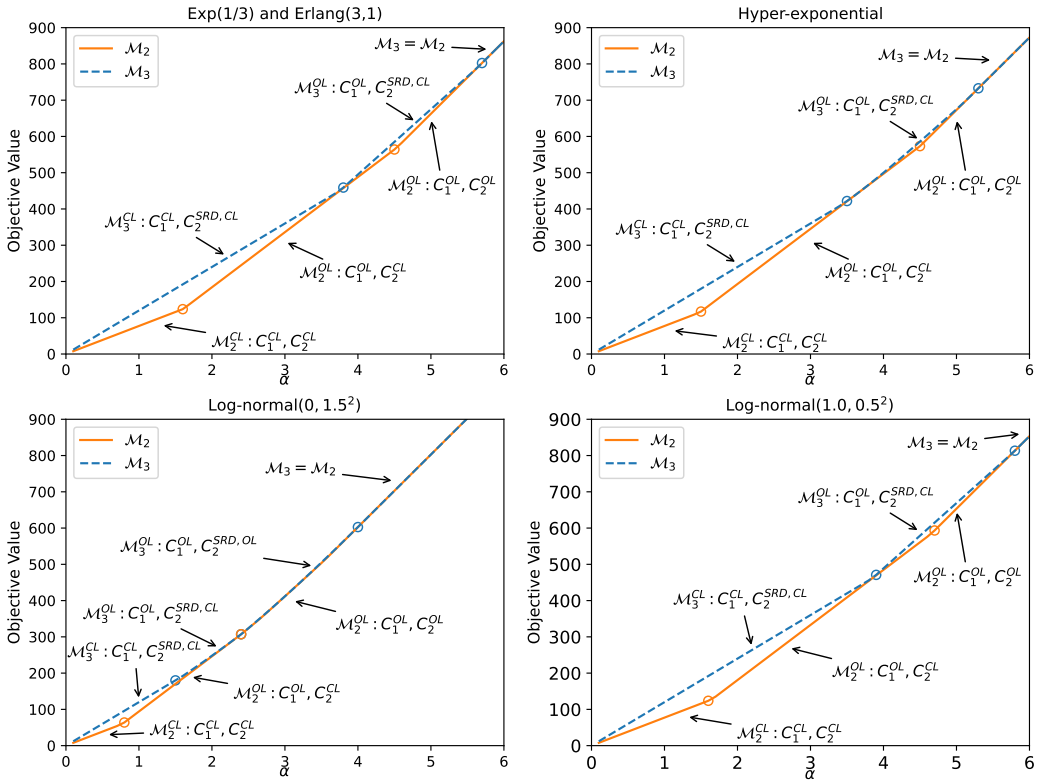
We next present a numerical example. Consider two customer classes with arrival rates $\Lambda_1 = \Lambda_2 = 75$ and $N = 100$ agents. System welfare is the total service reward received by served customers minus the cumulative waiting cost. The service reward function is common across the two classes and given by $V(\mu) = \alpha/\mu^2$, while the waiting-cost rates are class-specific, with $c_1 = 1$ and $c_2 = 1.5$. Following §4.3, we consider five patience time distributions: exponential, Erlang, hyper-exponential, log-normal$(0, 1.5^2)$, and log-normal$(1, 0.5^2)$. In this numerical example, both classes share the same patience time distribution. We vary α and, for each value of α, solve (21) and (22).

With two customer classes, Proposition 9 suggests that the optimal fluid solution uses at most three active service rate and offered wait pairs: one active pair for each class, plus at most one additional active pair created by splitting one of the classes. Figure 2 compares two values. The value $\mathcal{M}_3^*$ is the optimal objective value when one class is allowed to be split, so that up to three active service rate and offered wait pairs may be used. The value $\mathcal{M}_2^*$ is the optimal objective value under the restriction that each class uses a single service rate grade.

The figure also reports the operating regime and the class selected for service rate differentiation, when such differentiation is optimal. For example, the notation $\mathcal{M}_3^{\mathrm{OL}} : C_1^{\mathrm{OL}}, C_2^{\mathrm{SRD,CL}}$ denotes an overloaded system in which class 1 uses one service grade with a positive offered wait (OL), while class 2 is differentiated into two service rate grades (SRD), both with a zero offered wait (CL). Other labels can be interpreted analogously.

The numerical results show that within-class service rate differentiation can substantially improve system welfare even when customers are already heterogeneous across classes. In this example,

Figure 2: Optimal values of $\mathcal{M}_2^*$ and $\mathcal{M}_3^*$ under different patience time distributions.



the optimal policy differentiates class 2 but not class 1. The welfare improvement relative to the restricted benchmark $\mathcal{M}_2^*$ reaches 55.2% under exponential, Erlang, and hyper-exponential patience times, 56.2% under log-normal$(0, 1.5^2)$ patience times, and 55.0% under log-normal$(1, 0.5^2)$ patience times. These findings suggest that the main insight from the analysis of homogeneous customers extends qualitatively to heterogeneous customers: the value of service rate differentiation can be fully achieved by creating at most one additional service grade.

7 Conclusion

In this paper, we studied joint service rate and waiting time differentiation for homogeneous impatient customers in quality-based service systems. We developed a fluid formulation in which the manager chooses service grades, assigns customers to these grades, and schedules customers using their grade labels and elapsed waiting times. Our main structural result shows that, in the fluid model, there exists an optimal solution with at most two active service rate and offered wait pairs. Thus, the optimal policy can be reduced to one of the three forms: a single-grade FCFS policy, a single-grade policy with two offered waits, or a two-grade policy with FCFS within each grade. We further characterized the optimal policy under separable welfare functions. We show that both the shape of the service reward and patience time distribution can determine whether service rate differentiation should exist, and if yes, in what regimes. With strictly increasing or

constant hazard rates of patience times, any two-grade system, if optimal, must be critically loaded; overloaded systems, if optimal, can have only one grade. With strictly decreasing hazard rates of patience times, two-grade systems can be optimal in an overloaded regime, and a common offered wait is applied to both grades. We further propose an $\mathcal{S}$ policy to implement the fluid solution in stochastic systems through a randomized grade assignment at customer arrival and a priority rule based on customers' grade labels and elapsed waiting times. For Markovian systems, we provide a formal performance guarantee for this policy.

Several future directions are worth pursuing. First, the two-active-pair result is exact for the fluid model. However, an optimal policy in finite stochastic systems need not have the same form. Moreover, analyzing critically loaded systems may require additional refinements to accommodate stochastic fluctuations. We leave these refinements to future research. Second, extending the performance analysis of the differentiation policy to non-Markovian systems will further strengthen our understanding of this policy. Third, our results highlight the critical role of reward functions in implementing service rate differentiation. We focused on well-behaved reward functions commonly used in the quality-based service literature to develop insights, but reward functions in real-world systems may be nonsmooth or nonmonotone in service times. Estimating such reward functions in real contexts and integrating estimation with policy optimization is an important avenue for future research. Finally, we extended our analysis to heterogeneous customers, wherein the fluid solution suggests that at most one class needs to be further differentiated. A fuller analysis of service rate differentiation in multi-class systems, especially under class-specific fairness or service-level constraints, constitutes another relevant and interesting direction for future study.

Acknowledgments

C. Wu's research is generously supported by the Hong Kong Research Grants Council [Grant GRF 16501023]. W. You's research is generously supported by the Hong Kong Research Grants Council [Grant GRF 16212823] and [Theme-based Research Project T32-615/24-R].

References

Kranthi Mitra Adusumilli and John J Hasenbein. Dynamic admission and service rate control of a queue. *Queueing Systems*, 66(2):131–154, 2010.

Krishnan S Anand, M Fazıl Paç, and Senthil Veeraraghavan. Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Science*, 57(1):40–56, 2011.

Mor Armony and Avishai Mandelbaum. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Operations Research*, 59(1):50–65, 2011.

Mor Armony and Galit Bracha Yom-Tov. Hospital vs. home care: Trading off predischarge and postdischarge infection and mortality risks. *Manufacturing & Service Operations Management*, 28(1):57–75, 2026.

Barış Ata and Shiri Shneorson. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science*, 52(11):1778–1791, 2006.

Rami Atar, Chanit Giat, and Nahum Shimkin. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

Achal Bassamboo and Ramandeep S Randhawa. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research*, 58(5):1398–1413, 2010.

Achal Bassamboo and Ramandeep Singh Randhawa. Scheduling homogeneous impatient customers. *Management Science*, 62(7):2129–2147, 2016.

Achal Bassamboo, Ramandeep Randhawa, and Chenguang Wu. Optimally scheduling heterogeneous impatient customers. *Manufacturing & Service Operations Management*, 25(3):1066–1080, 2023.

Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

Jim G Dai, AB Dieker, and Xuefeng Gao. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems*, 78(1):1–29, 2014.

Jennifer M George and J Michael Harrison. Dynamic control of a queue with adjustable service rate. *Operations Research*, 49(5):720–731, 2001.

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

Wallace J Hopp, Seyed MR Iravani, and Gigi Y Yuen. Operations systems with discretionary task completion. *Management Science*, 53(1):61–77, 2007.

Kejia Hu, Gad Allon, and Achal Bassamboo. Understanding customer retrials in call centers: Preferences for service quality and service speed. *Manufacturing & Service Operations Management*, 24(2):1002–1020, 2022.

Ravi Kumar, Mark E Lewis, and Huseyin Topaloglu. Dynamic service rate control for a single-server queue with Markov-modulated arrivals. *Naval Research Logistics (NRL)*, 60(8):661–677, 2013.

Nelson Lee and Vidyadhar G Kulkarni. Optimal arrival rate and service rate control of multi-server queues. *Queueing Systems*, 76(1):37–50, 2014.

Gen Li, Jianhua Z Huang, and Haipeng Shen. To wait or not to wait: Two-way functional hazards model for understanding waiting in call centers. *Journal of the American Statistical Association*, 113(524): 1503–1514, 2018.

Zhenghua Long, Nahum Shimkin, Hailun Zhang, and Jiheng Zhang. Dynamic scheduling of multiclass many-server queues with abandonment: The generalized $c\mu/h$ rule. *Operations Research*, 68(4):1218–1230, 2020.

Jinting Wang, Zhongbin Wang, and Yunan Liu. Reducing delay in retrial queues by simultaneously differentiating service and retrial rates. *Operations Research*, 68(6):1648–1667, 2020.

Zhongbin Wang, Luyi Yang, Shiliang Cui, Sezer Ülkü, and Yong-Pin Zhou. Pooling agents for customer-intensive services. *Operations Research*, 71(3):860–875, 2023.

Ying Xu, Alan Scheller-Wolf, and Katia Sycara. The benefit of introducing variability in single-server queues with application to quality-based service domains. *Operations Research*, 63(1):233–246, 2015.

Sergey Zeltyn and Avishai Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems*, 51(3):361–402, 2005.

Dongyuan Zhan and Amy R Ward. Staffing, routing, and payment to trade off speed and quality in large service systems. *Operations Research*, 67(6):1738–1751, 2019.

Appendix

A An Equivalent View of the Policy

In this section, we provide an equivalent formulation of our grade allocation and scheduling problem, which entails an alternative interpretation and implementation of joint service rate and waiting time differentiation. Recall that our original policy creates grade differentiation in the first stage and uses the grade information to further create waiting time differentiation both across grades and within each grade in the second stage. Alternatively, one can first create waiting time differentiation using customers' elapsed waiting times and then create service rate differentiation among non-abandoning customers. Accordingly, in this alternative policy, the first stage focuses on assigning different offered waits to arriving customers, and the second stage optimizes the capacity and service rate(s) of each class, where classes are differentiated by their offered waits in the first stage.

Formally, with some abuse of notation, suppose the (homogeneous) arriving customers are assigned to one of finitely many offered waits in $\boldsymbol{w} = (w_1, w_2, \ldots, w_K)$ according to a probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$. The first-stage waiting time allocation problem can be stated as

$$\max_{\boldsymbol{w},\boldsymbol{p}} \ \mathcal{R}(\boldsymbol{w},\boldsymbol{p}),$$

with $\mathcal{R}(\boldsymbol{w},\boldsymbol{p})$ being the optimal value of the following capacity and service rate allocation problem:

$$\mathcal{R}(\boldsymbol{w},\boldsymbol{p}) = \max_{\boldsymbol{n}:\sum_i n_i \le N} \sum_i R(n_i|w_i,p_i),$$

where

$$\begin{aligned} R(n_i|w_i,p_i) = \sup_{\{J(i),n_{i,j},\mu_{i,j},\lambda_{i,j},j=1,\ldots,J(i)\}} \quad & \sum_j \lambda_{i,j} r(\mu_{i,j},w_i) \\ \text{s.t.} \quad & \sum_{j=1}^{J(i)} \lambda_{i,j} = \Lambda p_i, \\ & \lambda_{i,j}\bar{F}(w_i) \le n_{i,j}\mu_{i,j}, \\ & \sum_j n_{i,j} \le n_i. \end{aligned}$$

In the above, a customer class with arrival rate Λp_i and offered wait w_i is split into $J(i)$ subclasses. Each subclass represents a unique service rate, corresponding to our service rate differentiation (to be optimized jointly with the service capacity decisions), now implemented in the second stage. The following lemma shows that the new formulation is equivalent to the original problem in terms of system welfare.

Lemma 6. *For any feasible service-rate-first and waiting-time-next allocation* $(\boldsymbol{\mu},\boldsymbol{p},\boldsymbol{w})$*, there exists a feasible waiting-time-first and service rate-next allocation* $(\boldsymbol{w'},\boldsymbol{p'},\boldsymbol{\mu'})$ *with the same system welfare, and vice versa.*

The equivalence result in Lemma 6 above suggests that service rate differentiation can be implemented either upon customers' arrival or when they enter service. Such equivalence holds because abandoning customers don't enter service and service rates assigned to these customers upon arrival do not affect system performance. In other words, the benefits of service rate differentiation can only materialize among non-abandoning customers.

A.1 The $\mathcal{W}$ Policy

The $\mathcal{W}$ policy only creates service rate differentiation among non-abandoning customers when they enter service. Depending on the number of grades created in the fluid solution, the $\mathcal{W}$ policy can be stated as follows.

1. (Single-grade policy) If the fluid solution has only one grade μ^*, then process all customers either under FCFS (if the fluid solution has only one offered wait) or TIQ policy (if the fluid solution has two different offered waits) at rate μ^*.

2. (Two-grade policy) If the fluid solution has two grades $\mu_1^* \neq \mu_2^*$ with corresponding arrival rates $(\lambda_1^*, \lambda_2^*)$ and offered waits (w_1^*, w_2^*), then implement the following two steps.

   (a) If $w_1^* = w_2^*$, consider the following two cases.

   (a.1) If $w_1^* = w_2^* > 0$, then process all customers under FCFS.

   (a.2) If $w_1^* = w_2^* = 0$, then process all customers under LCFS.

   Process the next chosen customer at rate μ_1^* with probability λ_1^*/Λ and at rate μ_2^* with probability λ_2^*/Λ.

   (b) If $w_1^* \neq w_2^*$, suppose $w_1^* < w_2^*$ without loss of generality.

   (b.1) If $\bar{F}(w_1^*)/\mu_1^* \geq \bar{F}(w_2^*)/\mu_2^*$, then use the following modified TIQ policy.

   (i) First process customers who have waited more than w_2^* at rate μ_2^* under FCFS.

   (ii) Then process customers who have waited less than w_1^* at rate μ_1^* under FCFS.

   (iii) Next process the remaining customers at rate μ_1^* under LCFS.

   (b.2) Otherwise if $\bar{F}(w_1^*)/\mu_1^* < \bar{F}(w_2^*)/\mu_2^*$, then use the following policy.

   (i) First process customers with waiting times greater than w_1. With probability $\lambda_1^*\bar{F}(w_1^*)/(N\mu_1^*)$, process these customers under LCFS at rate μ_1^*, and with probability $\lambda_2^*\bar{F}(w_2^*)/(N\mu_2^*)$, process these customers under FCFS at rate μ_2^*.

   (ii) Next process the remaining customers at rate μ_2^* under FCFS.

The single-grade $\mathcal{W}$ policy replicates that of the $\mathcal{S}$ policy because service rate differentiation is moot. We next elaborate on the two-grade $\mathcal{W}$ policy.

When $w_1^* = w_2^*$, customers are not differentiated by their offered waits, leading to a natural choice of processing all customers under FCFS. Indeed, when $w_1^* = w_2^* > 0$, the system is overloaded and following FCFS is critical to achieving the desired offered wait. However, when $w_1^* = w_2^* = 0$, the system is critically loaded, and we advocate processing customers under LCFS. In fact, under

critical load, any non-idling policy can achieve a zero offered wait in the fluid model. Our choice of LCFS is based on its robust performance demonstrated in our numerical examples presented in §5.3, particularly notable under patience distributions with increasing hazard rates.

When $w_1^* \neq w_2^*$, the $\mathcal{W}$ policy has a simple implementation when $\bar{F}(w_1^*)/\mu_1^* \geq \bar{F}(w_2^*)/\mu_2^*$. Recall $w_1^* < w_2^*$ by our assumption, this condition states that the total workload of non-abandoning customers is decreasing in their waiting times. When this holds, we propose a modification (b.1) to the TIQ policy (Bassamboo and Randhawa 2016) to enable joint waiting time and service rate differentiation. To elaborate, note that $\Lambda\bar{F}(w_2^*)/\mu_2^* \leq N$ holds under this condition and thus step (i) ensures that in the steady state no fluid mass will wait for more than w_2^*. Similarly, we have $\Lambda\bar{F}(w_1^*)/\mu_1^* \geq N$ under this condition and thus step (ii) ensures that in the steady state no fluid mass will wait for less than w_1^*. These steps collectively ensure that customers in the system wait for either w_1^* or w_2^* time units before being served. Step (iii) plays a more subtle role. When $w_1^* > 0$, steps (i) and (ii) ensure that all capacities are exhausted in the fluid model and step (iii) is never applied. However, when $w_1^* = 0$, step (iii) is critical because step (ii) is never applied. In this case, LCFS in step (iii) ensures that all customers are either processed at rate μ_1^* when their time-in-queue is zero, or simply wait until their time-in-queue exceeds w_2^* and then are processed at rate μ_2^* when we apply step (i).

When $\bar{F}(w_1^*)/\mu_1^* < \bar{F}(w_2^*)/\mu_2^*$, the above modified TIQ policy can fail. Specifically, when we apply step (i), because $\Lambda\bar{F}(w_2^*)/\mu_2^* > N$, the system will likely stabilize at another offered wait $\tilde{w}_2 > w_2^*$ so that step (i) effectively processes all customers at rate μ_2^* under FCFS (with offered wait $\tilde{w}_2$). Thus, the policy may fail to lead us to the desired two-grade fluid solution. In this case, we propose prioritizing customers with waiting times greater than w_1^*. (Because if we apply step (ii) in the modified TIQ policy (b.1), the system may, likewise, stabilize at another offered wait $\tilde{w}_1 < w_1^*$.) When we apply step (i) of (b.2), because there is insufficient capacity to process all customers with waiting times greater than w_1^* under LCFS, those unprocessed will wait until their time-in-queue exceeds w_2^* and then be processed under FCFS. For customers with waiting times less than w_1^*, we propose step (ii) to ensure non-idling of servers, and because $\Lambda\bar{F}(w_1^*)/\mu_1^* > \Lambda\bar{F}(w_2^*)/\mu_2^* > N$, this step also ensures that in the steady state no fluid mass will wait less than w_1^*.

A.2 Performance of the $\mathcal{W}$ Policy for Markovian Systems

The following result mirrors Proposition 8, showing that the $\mathcal{W}$ policy shares the same performance guarantee as the $\mathcal{S}$ policy. Let $K_\Lambda^{\mathcal{W}}$ denote the steady-state system welfare under the $\mathcal{W}$ policy.

Proposition 10. *Suppose the inter-arrival times, service times and patience times are exponentially distributed, and suppose that the welfare function is separable. Then the $\mathcal{W}$ policy has $\mathcal{O}(\sqrt{\Lambda})$ performance. That is, there exists finite $A > 0$ such that*

$$K_\Lambda^* - K_\Lambda^{\mathcal{W}} \leq A\sqrt{\Lambda} \qquad \text{as } \Lambda \to \infty.$$

Moreover, if the optimal value $\mathcal{M}^*(N, \Lambda)$ is achieved under (11) with $w^* > 0$, then the $\mathcal{W}$ policy

has $o(1)$ *performance:*

$$\left|K_\Lambda^* - K_\Lambda^{\mathcal{W}}\right| \to 0 \qquad \text{as } \Lambda \to \infty.$$

Further, the optimal stochastic-system welfare is upper bounded by the optimal objective value of the fluid optimization problems (11) *and* (13)*. That is,* $K_\Lambda^* \le \mathcal{M}^*(N, \Lambda)$.

A.3 Comparison between $\mathcal{S}$ and $\mathcal{W}$ policies.

The $\mathcal{S}$ and $\mathcal{W}$ policies implement the same fluid solution but differ in when service rate differentiation is created. The $\mathcal{S}$ policy assigns each customer a grade upon arrival and then uses this grade information for scheduling. This policy requires the system to store each customer's grade information while the customer is waiting. In contrast, the $\mathcal{W}$ policy assigns service rates only to non-abandoning customers when they enter service. This reduces implementation complexity.

Both policies have the same fluid-level performance. In finite stochastic systems, however, their performance can slightly differ. Table 5 compares the performance of these two policies via simulations. The $\mathcal{S}$ policy performs slightly better in most cases, consistent with the fact that two-grade $\mathcal{S}$ policies can exploit service time heterogeneity in scheduling; see Xu et al. (2015) for a related observation. The welfare differences between the two policies are small in most cases, so the choice between these two policies is primarily driven by an implementation tradeoff: $\mathcal{W}$ policy is preferable when simplicity is important, whereas $\mathcal{S}$ policy is preferable when robustness of performance is important.

Table 5 also reports a variant of the $\mathcal{W}$ policy that uses FCFS, rather than LCFS, in step (a.2) for two-grade critically loaded systems; see the "CL-FCFS" columns. This variant can suffer substantial losses under Erlang and some log-normal patience distributions, for which LCFS minimizes the within-class queue length, while it yields only minor gains under the hyper-exponential distribution, for which FCFS minimizes the within-class queue length. We therefore recommend LCFS to be used in two-grade critically loaded systems as a more robust policy.

Table 5: Comparison of $\mathcal{S}$ and $\mathcal{W}$ policies: simulation, $V(\mu) = \alpha/\mu^2$, $\Lambda = 150$, $N = 100$, and $\mu \in [0.5, 2.5]$.

	$\alpha = 2$			$\alpha = 3$		
Patience Distribution	$\mathcal{S}$ policy	$\mathcal{W}$ policy	$\mathcal{W}$ policy (CL-FCFS)	$\mathcal{S}$ policy	$\mathcal{W}$ policy	$\mathcal{W}$ policy (CL-FCFS)
Exp(1/3)	221.1	$-0.5\%^*$	$-0.5\%^*$	335.0	$-0.4\%^*$	$-0.5\%^*$
Erlang(3,1)	218.2	$-0.8\%^*$	$-14.5\%^*$	334.6	$-0.4\%^*$	$-8.7\%^*$
Hyper-exponential	221.3	$-0.4\%^*$	-0.3%	335.1	$+0.5\%^*$	$+0.5\%^*$
Log-normal$(0, 1.5^2)$	255.4	-0.0%	-0.0%	439.3	-0.2%	-0.2%
Log-normal$(1, 0.5^2)$	215.3	$+0.2\%$	$-25.7\%^*$	332.5	$+0.0\%$	$-16.0\%^*$

*Statistically significant at the 5% level.

Entries in the $\mathcal{W}$ columns report percentage changes relative to the corresponding $\mathcal{S}$ policy.

B Proofs of Main Results

Proof. Proof of Proposition 1. The problems (6), (7), and (10) can be reformulated as

$$\begin{aligned} \sup_{\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{w}} \quad & \sum_{i,j} \lambda_{i,j} r(\mu_i, w_{i,j}) \\ \text{s.t.} \quad & \sum_{i,j} \lambda_{i,j} = \Lambda, \\ & \sum_{i,j} \lambda_{i,j} \bar{F}(w_{i,j})/\mu_i \leq N. \end{aligned} \tag{24}$$

To see this, substitute (7) into (10) and let $\boldsymbol{w}_i = (w_{i,j} : j = 1, 2, \ldots, J(i))$, $\boldsymbol{\lambda}_i = (\lambda_{i,j} : j = 1, 2, \ldots, J(i))$, and $\boldsymbol{n}_i = (n_{i,j} : j = 1, 2, \ldots, J(i))$. We have

$$\begin{aligned} \max_{\boldsymbol{\mu},\boldsymbol{p}} \ \mathcal{R}(\boldsymbol{\mu}, \boldsymbol{p}) \ = \max_{\boldsymbol{\mu},\boldsymbol{p}} \sup_{\boldsymbol{w}_i,\boldsymbol{\lambda}_i,\boldsymbol{n}_i} \quad & \sum_{i,j} \lambda_{i,j} r(\mu_i, w_{i,j}) \\ \text{s.t.} \quad & \sum_{j} \lambda_{i,j} = \Lambda p_i, \\ & \lambda_{i,j} \bar{F}(w_{i,j})/\mu_i \leq n_{i,j}, \\ & \sum_{j} n_{i,j} = n_i, \\ & \sum_{i} n_i \leq N. \end{aligned} \tag{25}$$

The new objective function is the same as (24). Regarding the constraints, note that any feasible solution of (25) is a feasible solution of (24). Likewise, for any feasible solution of (24), we can use it to construct a feasible solution of (25) by setting $n_{i,j} = \lambda_{i,j} \bar{F}(w_{i,j})/\mu_i$ and $p_i = \sum_j \lambda_{i,j}/\Lambda$. This establishes the equivalence between the two problems. We next focus on (24) to complete our analysis.

Assume that the supremum in (24) is achieved for some K^*, J_i^*, $\lambda_{i,j}^*$, and $w_{i,j}^*$. Then, (24) has an optimal objective value identical to that of the following linear program for $\boldsymbol{\lambda}$:

$$\begin{aligned} \mathcal{R}(\boldsymbol{\mu}) \ = \ \sup_{\boldsymbol{\lambda}} \quad & \sum_{i,j} \lambda_{i,j} r(\mu_i, w_{i,j}^*) \\ \text{s.t.} \quad & \sum_{i,j} \lambda_{i,j} = \Lambda, \\ & \sum_{i,j} \lambda_{i,j} \bar{F}(w_{i,j}^*)/\mu_i \leq N. \end{aligned} \tag{26}$$

Standard linear programming theory implies there exists an optimal solution to the linear program (26) that has at most 2 positive components. This follows because there exists an optimal solution that lies in the set of basic solutions, and further, because there are 2 constraints, the basis can have a rank of at most 2. Thus, any basic feasible solution can have at most 2 positive components. □

Proof. Proof of Proposition 2. We present the proof for the two-grade optimization (13). The proof for the single-grade problem (11) follows similarly.

By solving λ_1 and λ_2 as functions of $\{\mu_1, \mu_2, w_1, w_2\}$, the two-grade problem (13) can be reformulated as

$$\begin{aligned} \max_{\mu_1,\mu_2,w_1,w_2} \quad & G(\mu_1, \mu_2, w_1, w_2) \\ \text{s.t.} \quad & \mu_2 \geq \underline{\mu} \\ & \mu_1 \geq \mu_2 \\ & -\mu_1 \geq -\bar{\mu} \\ & w_1 \geq 0 \\ & w_2 \geq 0, \end{aligned}$$

where $G(\mu_1, \mu_2, w_1, w_2) = \lambda_1 r(\mu_1, w_1) + \lambda_2 r(\mu_2, w_2)$ with λ_1 and λ_2 defined in (14). The Lagrangian can be written as

$$\mathcal{L}(\mu_1, \mu_2, w_1, w_2) = G(\mu_1, \mu_2, w_1, w_2) + \xi_1(\mu_2 - \underline{\mu}) + \xi_2(\mu_1 - \mu_2) + \xi_3(\bar{\mu} - \mu_1) + \xi_4 w_1 + \xi_5 w_2,$$

where ξ_i's are the Lagrangian multipliers corresponding to five constraints. The following partial derivatives hold

$$\begin{aligned} \frac{\partial \lambda_1}{\partial w_1} &= -\frac{\partial \lambda_2}{\partial w_1} = \frac{\lambda_1 f(w_1)/\mu_1}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}, \\ \frac{\partial \lambda_1}{\partial w_2} &= -\frac{\partial \lambda_2}{\partial w_2} = \frac{\lambda_2 f(w_2)/\mu_2}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}, \\ \frac{\partial G}{\partial w_1} &= \frac{\lambda_1 f(w_1)/\mu_1}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2} \left(r(\mu_1, w_1) - r(\mu_2, w_2)\right) + \lambda_1 \frac{\partial r(\mu_1, w_1)}{\partial w_1}, \\ \frac{\partial G}{\partial w_2} &= \frac{\lambda_2 f(w_2)/\mu_2}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2} \left(r(\mu_1, w_1) - r(\mu_2, w_2)\right) + \lambda_2 \frac{\partial r(\mu_2, w_2)}{\partial w_2}. \end{aligned}$$

When an optimal solution w_1 satisfies $0 < w_1 < \infty$, complementary slackness suggests $\xi_4 = 0$ and thus

$$\frac{\partial \mathcal{L}}{\partial w_1} = 0 \quad \Longleftrightarrow \quad \frac{\partial G}{\partial w_1} = 0 \quad \Longleftrightarrow \quad \frac{\frac{\partial r(\mu_1, w_1)}{\partial w_1} \cdot \mu_1}{f(w_1)} = -\frac{r(\mu_1, w_1) - r(\mu_2, w_2)}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}. \tag{27}$$

When an optimal solution w_2 satisfies $0 < w_2 < \infty$, complementary slackness suggests $\xi_5 = 0$ and thus

$$\frac{\partial \mathcal{L}}{\partial w_2} = 0 \quad \Longleftrightarrow \quad \frac{\partial G}{\partial w_2} = 0 \quad \Longleftrightarrow \quad \frac{\frac{\partial r(\mu_2, w_2)}{\partial w_2} \cdot \mu_2}{f(w_2)} = -\frac{r(\mu_1, w_1) - r(\mu_2, w_2)}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}. \tag{28}$$

Combining (27) and (28), we have the desired result. □

Proof. Proof of Proposition 3.

We first show Part 1 of Proposition 3. Consider first the single-grade problem (11). Plugging in the solution of λ_l and λ_h defined in (12), we rewrite the objective function as

$$\lambda_l r(\mu, w_l) + \lambda_h r(\mu, w_h) = NV(\mu)\mu - \sum_{j=l,h} \lambda_j \int_0^{w_j} \bar{F}(y) dy.$$

Since $V(\mu)\mu$ is weakly increasing, it is optimal to set $\mu = \mu_0$, leading to $w_l = w_h = 0$. Next, for the two-grade problem, we have

$$r(\mu_i, w_i) = \min\{n_i\mu_i, \lambda_i\}V(\mu_i) - \lambda_i \int_0^{w_i} \bar{F}(y)dy, \quad i = 1, 2.$$

Suppose we have a feasible solution $(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{n}, \boldsymbol{w})$ such that $\boldsymbol{w} \neq 0$. For any i, $w_i > 0$ implies $n_i\mu_i > \lambda_i$. Since $V(\mu)\mu$ is weakly increasing, we can always increase μ_i to λ_i/n_i (which results in $w_i = 0$) and achieve an objective value at least as good as before. This implies that there always exists an optimal solution with $\boldsymbol{w} = 0$, leading to a critically-loaded system.

Next, we show Part 2 of Proposition 3. Suppose we have a feasible solution with $\mu_1 > \mu_2$. We will construct another feasible solution with $\mu_1 = \mu_2 = \tilde{\mu}$ that achieves an equal or higher objective value. To this end, fix the values of $\lambda_1, \lambda_2, w_1$, and w_2. The capacity constraint holds

$$\frac{\lambda_1\bar{F}(w_1)}{N\mu_1} + \frac{\lambda_2\bar{F}(w_2)}{N\mu_2} = 1.$$

Define $\tau_1 = \frac{\lambda_1\bar{F}(w_1)}{N\mu_1}$ and $\tau_2 = \frac{\lambda_2\bar{F}(w_2)}{N\mu_2}$. Then, the objective function can be expressed as

$$N(\tau_1 V(\mu_1)\mu_1 + \tau_2 V(\mu_2)\mu_2) - \lambda_1 \int_0^{w_1} \bar{F}(y)dy - \lambda_2 \int_0^{w_2} \bar{F}(y)dy.$$

The concavity of $V(\mu)\mu$ and Jensen's inequality further imply that

$$\tau_1 V(\mu_1)\mu_1 + \tau_2 V(\mu_2)\mu_2 \leq V(\tilde{\mu})\tilde{\mu},$$

where $\tilde{\mu} = \tau_1\mu_1 + \tau_2\mu_2$ constitutes a feasible solution satisfying the capacity constraint. Now, the objective value of the single-grade problem with $\tilde{\mu}$, $\boldsymbol{\lambda}$, and $\boldsymbol{w}$ is

$$NV(\tilde{\mu})\tilde{\mu} - \lambda_1 \int_0^{w_1} \bar{F}(y)dy - \lambda_2 \int_0^{w_2} \bar{F}(y)dy.$$

Thus, we have an alternative single-grade solution that achieves a weakly higher objective value. □

Proof. Proof of Lemma 1. Consider the two-grade problem. Using the notation from the proof of Proposition 2, we have the following partial derivatives

$$\begin{aligned}
\frac{\partial G}{\partial \mu_1} &= \frac{\lambda_1\bar{F}(w_1)}{\mu_1^2}\left(D + V'(\mu_1)\mu_1^2\right) \\
\frac{\partial G}{\partial \mu_2} &= \frac{\lambda_2\bar{F}(w_2)}{\mu_2^2}\left(D + V'(\mu_2)\mu_2^2\right) \\
\frac{\partial G}{\partial w_1} &= \frac{\lambda_1 f(w_1)}{\mu_1}\left(D - V(\mu_1)\mu_1 - \frac{\mu_1}{H(w_1)}\right) \\
\frac{\partial G}{\partial w_2} &= \frac{\lambda_2 f(w_2)}{\mu_2}\left(D - V(\mu_2)\mu_2 - \frac{\mu_2}{H(w_2)}\right),
\end{aligned} \tag{29}$$

where

$$D \equiv D(\mu_1, \mu_2, w_1, w_2) = \frac{V(\mu_1)\bar{F}(w_1) - V(\mu_2)\bar{F}(w_2) + \int_{w_1}^{w_2} \bar{F}(y)dy}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}.$$

Plugging in $w_1 = w_2 = 0$, we have

$$D(\mu_1, \mu_2, 0, 0) = \frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2}.$$

When an optimal solution (μ_1, μ_2) satisfies $\underline{\mu} < \mu_2 < \mu_1 < \bar{\mu}$, complementary slackness suggests $\xi_1 = \xi_2 = \xi_3 = 0$ and thus

$$\frac{\partial \mathcal{L}}{\partial w_1} = 0, \frac{\partial \mathcal{L}}{\partial w_2} = 0 \quad \Longleftrightarrow \quad V'(\mu_1) \cdot \mu_1^2 = V'(\mu_2) \cdot \mu_2^2 = -\frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2}.$$

When the optimal (μ_1, μ_2) has the form of $(\mu_1, \underline{\mu})$ with $\mu_2 < \mu_1 < \bar{\mu}$, then $\xi_2 = \xi_3 = 0$ and

$$\frac{\partial \mathcal{L}}{\partial w_1} = 0 \quad \Longleftrightarrow \quad V'(\mu_1) \cdot \mu_1^2 = -\frac{V(\mu_1) - V(\underline{\mu})}{1/\mu_1 - 1/\underline{\mu}}.$$

When the optimal (μ_1, μ_2) has the form of $(\bar{\mu}, \mu_2)$ with $\underline{\mu} < \mu_2 < \mu_1$, then $\xi_1 = \xi_2 = 0$ and

$$\frac{\partial \mathcal{L}}{\partial w_2} = 0 \quad \Longleftrightarrow \quad V'(\mu_2) \cdot \mu_2^2 = -\frac{V(\bar{\mu}) - V(\mu_2)}{1/\bar{\mu} - 1/\mu_2}.$$

Combining these, we have the desired result. □

Proof. Proof of Proposition 4. We first show that none of the first-order conditions in Lemma 1 can hold when $-V'(\mu) \cdot \mu^2$ is strictly monotone. For any $\mu_1 > \mu_2$, note that

$$\frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2} = \frac{\int_{\mu_2}^{\mu_1} [-V'(\mu)]\, d\mu}{\int_{\mu_2}^{\mu_1} \mu^{-2}\, d\mu} = \frac{\int_{\mu_2}^{\mu_1} [-V'(\mu)\mu^2]\mu^{-2}\, d\mu}{\int_{\mu_2}^{\mu_1} \mu^{-2}\, d\mu}.$$

Hence, the right-hand side of the first-order conditions in Lemma 1 is a weighted average of $-V'(\mu)\mu^2$ over $[\mu_2, \mu_1]$.

Suppose first that $-V'(\mu)\mu^2$ is strictly decreasing. Since $\mu_1 > \mu_2$, we have

$$-V'(\mu_1)\mu_1^2 < \frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2} < -V'(\mu_2)\mu_2^2.$$

Hence, Part 1 of Lemma 1 cannot hold. The same weighted-average argument shows that Parts 2 and 3 of Lemma 1 cannot hold either. Therefore, no interior or one-sided boundary solution satisfying the necessary first-order conditions can be optimal.

Using the partial derivatives in (29) with $(w_1, w_2) = (0, 0)$, we have

$$\frac{\partial G}{\partial \mu_1} = \frac{\lambda_1}{\mu_1^2}\left[\frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2} - [-V'(\mu_1)\mu_1^2]\right] > 0$$

for all $\mu_1 < \bar{\mu}$, and

$$\frac{\partial G}{\partial \mu_2} = \frac{\lambda_2}{\mu_2^2}\left[\frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2} - [-V'(\mu_2)\mu_2^2]\right] < 0$$

for all $\mu_2 > \underline{\mu}$. Combining these two inequalities, the optimal solution must be $(\bar{\mu}, \underline{\mu})$ when $-V'(\mu)\mu^2$ is strictly decreasing.

Next, suppose that $-V'(\mu)\mu^2$ is strictly increasing. The same weighted-average argument gives

$$-V'(\mu_1)\mu_1^2 > \frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2} > -V'(\mu_2)\mu_2^2.$$

Therefore,

$$\frac{\partial G}{\partial \mu_1} < 0 \quad \text{and} \quad \frac{\partial G}{\partial \mu_2} > 0.$$

Thus, the objective improves by moving μ_1 downward and μ_2 upward. Since feasibility of the critically loaded two-grade problem requires $\mu_2 \le \mu_0 \le \mu_1$, the optimal value is obtained by collapsing the two service rates to the single critically loaded rate μ_0. Therefore, $\mathcal{M}_2^{\mathrm{CL}} = \mathcal{M}_1^{\mathrm{CL}}$. □

Proof. Proof of Lemma 2.

Following the notation in (29) and Proposition 2, we write the optimality condition for a solution (w_1, w_2) with $0 < w_1, w_2 < \infty$ as

$$V(\mu_1)\mu_1 + \frac{\mu_1}{H(w_1)} = V(\mu_2)\mu_2 + \frac{\mu_2}{H(w_2)} = D, \tag{30}$$

where

$$D = \frac{V(\mu_1)\bar{F}(w_1) - V(\mu_2)\bar{F}(w_2) + \int_{w_1}^{w_2} \bar{F}(y)dy}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}.$$

We next show $D - V(\mu_1)\mu_1 - \frac{\mu_1}{H(w_1)}$ cannot be zero, thus violating (30). To this end, we focus on analyzing the sign of $D - V(\mu_1)\mu_1 - \frac{\mu_1}{H(w_1)}$, which can be expressed as

$$\begin{aligned}
&D - V_1\mu_1 - \frac{\mu_1}{H(w_1)} \\
&= \frac{V_1\bar{F}(w_1) - V_2\bar{F}(w_2) + \int_{w_1}^{w_2} \bar{F}(y)dy - V_1\bar{F}(w_1) + V_1\bar{F}(w_2)\frac{\mu_1}{\mu_2} - \frac{1}{H(w_1)}\left(\bar{F}(w_1) - \bar{F}(w_2)\frac{\mu_1}{\mu_2}\right)}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2} \\
&= \frac{\frac{\bar{F}(w_2)}{\mu_2}(V_1\mu_1 + \frac{\mu_1}{H(w_1)} - V_2\mu_2) + \int_{w_1}^{w_2} \bar{F}(y)dy - \frac{\bar{F}(w_1)}{H(w_1)}}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2} \\
&= \frac{\frac{\bar{F}(w_2)}{H(w_2)} - \frac{\bar{F}(w_1)}{H(w_1)} + \int_{w_1}^{w_2} \bar{F}(y)dy}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2},
\end{aligned} \tag{31}$$

where the last equality follows from the first equality in the optimality condition (30).

Now, suppose the hazard rate of the patience time distribution is strictly decreasing.

- When $w_1 > w_2$ (possibly $w_1 = \infty$), we have $H(w_1) < H(y) < H(w_2)$ for $y \in (w_2, w_1)$, which implies $\frac{f(y)}{H(w_2)} < \bar{F}(y) < \frac{f(y)}{H(w_1)}$. Since $w_1 > w_2$, we have

  $$\frac{\bar{F}(w_1) - \bar{F}(w_2)}{H(w_1)} < \int_{w_1}^{w_2} \bar{F}(y)dy < \frac{\bar{F}(w_1) - \bar{F}(w_2)}{H(w_2)}.$$

  Thus,

  $$\bar{F}(w_2)\left(\frac{1}{H(w_2)} - \frac{1}{H(w_1)}\right) < \frac{\bar{F}(w_2)}{H(w_2)} - \frac{\bar{F}(w_1)}{H(w_1)} + \int_{w_1}^{w_2} \bar{F}(y)dy < \bar{F}(w_1)\left(\frac{1}{H(w_2)} - \frac{1}{H(w_1)}\right).$$

  Note that the lower and upper bound in the inequalities above have the same sign, which implies the numerator of (31) cannot be 0. Hence, $D - V_1\mu_1 - \frac{\mu_1}{H(w_1)} = 0$ cannot hold.

- When $w_1 < w_2$ (possibly $w_2 = \infty$), we have $H(w_2) < H(y) < H(w_1)$ for $y \in (w_1, w_2)$. In this case, we have

$$\bar{F}(w_2)\left(\frac{1}{H(w_2)} - \frac{1}{H(w_1)}\right) < \frac{\bar{F}(w_2)}{H(w_2)} - \frac{\bar{F}(w_1)}{H(w_1)} + \int_{w_1}^{w_2} \bar{F}(y)dy < \bar{F}(w_1)\left(\frac{1}{H(w_2)} - \frac{1}{H(w_1)}\right).$$

  Similarly, the lower and upper bound in the inequalities above have the same sign, which implies the numerator of (31) cannot be 0. Hence, $D - V_1\mu_1 - \frac{\mu_1}{H(w_1)} = 0$ cannot hold.

The case of patience time distributions with an increasing hazard rate follows analogously. We omit the proof for brevity. □

Proof. Proof of Lemma 3. In the case of increasing hazard rates, the optimal offered waits for each grade are $(0, \infty)$, i.e., to serve that grade under LCFS. In particular, for a feasible solution with a finite and positive w_1 or w_2, we should always split that grade into two subgrades. This is because the marginal value of increasing capacity allocated to a subgrade with a shorter offered wait is always greater than the marginal cost of reducing capacity allocated to the other subgrade with a longer offered wait. Thus, to optimally process a grade (with insufficient capacity to process it all), we should split that grade into two subgrades, one with offered wait zero and the other with offered wait infinity. □

Proof. Proof of Proposition 7. To simplify notation, we write $V_i := V(\mu_i)$ for $i = 1, 2$.

Case 1: Suppose $V_1\mu_1 + \mu_1/\gamma \geq V_2\mu_2 + \mu_2/\gamma$. In this case, grade 1 will be prioritized. Under fixed $\boldsymbol{\mu}$, we optimize over λ_1 and write the total welfare $R(\lambda_1)$ as a function of λ_1. Define $\tilde{\Lambda}_1 := \frac{\Lambda - N\mu_2}{1-\mu_2/\mu_1}$.

Case 1.1: $\mu_1 > \mu_0$. The total welfare

$$R(\lambda_1) = \begin{cases} V_1\lambda_1 + V_2(N - \lambda_1/\mu_1)\mu_2 - \left[\Lambda - \lambda_1 - (N - \lambda_1/\mu_1)\mu_2\right]/\gamma & \text{for } \lambda_1 \leq \tilde{\Lambda}_1, \\ V_1\lambda_1 + V_2(\Lambda - \lambda_1) & \text{for } \tilde{\Lambda}_1 < \lambda_1 \leq \Lambda. \end{cases}$$

Thus, $R'(\lambda_1) = V_1 - V_2 \cdot \mu_2/\mu_1 + (1 - \mu_2/\mu_1)/\gamma \geq 0$ for $\lambda_1 \leq \tilde{\Lambda}_1$ and $R'(\lambda_1) = V_1 - V_2 < 0$ for $\tilde{\Lambda}_1 < \lambda_1 \leq \Lambda$. Hence, the optimal $\lambda_1^* = \tilde{\Lambda}_1$. This leads to $(w_1, w_2) = (0, 0)$ and recovers (15).

Case 1.2: $\mu_1 \leq \mu_0$. The total welfare

$$R(\lambda_1) = \begin{cases} V_1\lambda_1 + V_2(N - \lambda_1/\mu_1)\mu_2 - \left[\Lambda - \lambda_1 - (N - \lambda_1/\mu_1)\mu_2\right]/\gamma & \text{for } \lambda_1 \leq N\mu_1, \\ V_1 N\mu_1 - (\Lambda - N\mu_1)/\gamma & \text{for } N\mu_1 < \lambda_1 \leq \Lambda. \end{cases}$$

Thus, the optimal $\lambda_1^* = N\mu_1$. This leads to $(w_1, w_2) = (0, \infty)$ and recovers $\mathcal{M}_1^{\text{LCFS}}$.

Case 2: Suppose $V_1\mu_1 + \mu_1/\gamma < V_2\mu_2 + \mu_2/\gamma$. Grade 2 will be prioritized. Under fixed $\boldsymbol{\mu}$, we optimize over λ_2 and write the total welfare $R(\lambda_2)$ as a function of λ_2. Define $\tilde{\Lambda}_2 = \frac{N\mu_1 - N}{\mu_1/\mu_2 - 1}$.

Case 2.1: $\mu_1 > \mu_0$. The total welfare

$$R(\lambda_2) = \begin{cases} V_1(\Lambda - \lambda_2) + V_2\lambda_2 & \text{for } 0 < \lambda_2 \leq \tilde{\Lambda}_2, \\ V_2\lambda_2 + V_1(N - \lambda_2/\mu_2)\mu_1 - \left[\Lambda - \lambda_2 - (N - \lambda_2/\mu_2)\mu_1\right]/\gamma & \text{for } \tilde{\Lambda}_2 \leq \lambda_2 \leq N\mu_2, \\ V_2 N\mu_2 - (\Lambda - N\mu_2)/\gamma & \text{for } N\mu_2 < \lambda_2 \leq \Lambda. \end{cases}$$

Thus, the optimal $\lambda_2^* = N\mu_2$. This leads to $(w_1, w_2) = (\infty, 0)$ and recovers $\mathcal{M}_1^{\text{LCFS}}$.

Case 2.2: $\mu_1 \le \mu_0$. The total welfare

$$R(\lambda_2) = \begin{cases} V_2\lambda_2 + V_1(N - \lambda_2/\mu_2)\mu_1 - [\Lambda - \lambda_2 - (N - \lambda_2/\mu_2)\mu_1]/\gamma & \text{for } 0 < \lambda_2 \le N\mu_2, \\ V_2 N\mu_2 - (\Lambda - N\mu_2)/\gamma & \text{for } N\mu_2 < \lambda_2 \le \Lambda. \end{cases}$$

Thus, the optimal $\lambda_2^* = N\mu_2$. This leads to $(w_1, w_2) = (\infty, 0)$ and recovers $\mathcal{M}_1^{\text{LCFS}}$.

Combining all cases, we have the desired result.

□

Proof. Proof of Corollary 2. Recall that the optimization problem under a single-grade policy is

$$\mu^* = \underset{\underline{\mu} \le \mu \le \mu_0}{\text{argmax}} \; NV(\mu) \cdot \mu - (\Lambda - N\mu)/\gamma = \underset{\underline{\mu} \le \mu \le \mu_0}{\text{argmax}} \, \mu[V(\mu) + 1/\gamma].$$

When $\mu\,[V(\mu) + 1/\gamma]$ is strictly increasing, we have $\mu^* = \mu_0$ and $\mathcal{M}_1^* = M_1(\mu_0)$. By Proposition 5, the optimal welfare is the larger one between $M_1(\mu_0)$ and M_2^{CL} in (15).

When $\mu\,[V(\mu) + 1/\gamma]$ is strictly decreasing, we have $\mu^* = \underline{\mu}$ and $\mathcal{M}_1^* = M_1(\underline{\mu})$ in the single-grade problem. If instead, two-grade service rates $\boldsymbol{\mu}^*$ are optimal, then following the proof of Proposition 7, they must satisfy $\mu_1^*[V(\mu_1^*) + 1/\gamma] \ge \mu_2^*[V(\mu_2^*) + 1/\gamma]$, i.e., the faster grade is prioritized. However, this condition is not valid when $\mu[V(\mu) + 1/\gamma]$ is strictly decreasing. Hence, the optimal policy should have only one grade $\underline{\mu}$. □

Proof of Lemma 4. Consider first $w_1 > 0$ and $w_2 = 0$. The first-order conditions are

$$\begin{aligned} D - V_1\mu_1 - \frac{\mu_1}{H(w_1)} &= 0, \\ D - V_2\mu_2 - \frac{\mu_2}{H(w_2)} &= -\xi_5 \frac{\mu_2}{\lambda_2 f(w_2)}. \end{aligned}$$

where $\xi_5 \ge 0$ is the Lagragian multiplier corresponding to the constraint $w_2 \ge 0$ (see the proof of Proposition 2) and

$$D = \frac{V(\mu_1)\bar{F}(w_1) - V(\mu_2)\bar{F}(w_2) + \int_{w_1}^{w_2} \bar{F}(y)dy}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}.$$

Thus,

$$V_1\mu_1 + \frac{\mu_1}{H(w_1)} - V_2\mu_2 = \frac{\mu_2}{H(w_2)} - \xi_5 \frac{\mu_2}{\lambda_2 f(w_2)}. \tag{32}$$

We now show $D - V(\mu_1)\mu_1 - \frac{\mu_1}{H(w_1)}$ cannot be zero, thus violating the optimality condition for w_1. To this end, we focus on analyzing the sign of $D - V(\mu_1)\mu_1 - \frac{\mu_1}{H(w_1)}$. Similar to (31), we have

$$\begin{aligned} D - V_1\mu_1 - \frac{\mu_1}{H(w_1)} &= \frac{\frac{\bar{F}(w_2)}{\mu_2}(V_1\mu_1 + \frac{\mu_1}{H(w_1)} - V_2\mu_2) + \int_{w_1}^{w_2} \bar{F}(y)dy - \frac{\bar{F}(w_1)}{H(w_1)}}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2} \\ &= \frac{\frac{\bar{F}(w_2)}{\mu_2}(\frac{\mu_2}{H(w_2)} - \xi_5\frac{\mu_2}{\lambda_2 f(w_2)}) + \int_{w_1}^{w_2} \bar{F}(y)dy - \frac{\bar{F}(w_1)}{H(w_1)}}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2} \\ &= \frac{-\frac{\xi_5}{\lambda_2 H(w_2)} + \frac{\bar{F}(w_2)}{H(w_2)} - \frac{\bar{F}(w_1)}{H(w_1)} + \int_{w_1}^{w_2} \bar{F}(y)dy}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}, \end{aligned}$$

where the second equality follows from (32).

Since $H(\cdot)$ is decreasing and $w_1 > w_2$, we have

$$\bar{F}(w_2)\left(\frac{1}{H(w_2)}-\frac{1}{H(w_1)}\right) < \frac{\bar{F}(w_2)}{H(w_2)} - \frac{\bar{F}(w_1)}{H(w_1)} + \int_{w_1}^{w_2} \bar{F}(y)dy < \bar{F}(w_1)\left(\frac{1}{H(w_2)}-\frac{1}{H(w_1)}\right).$$

Now, since $\mu_1 > \mu_2$, we have $\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2 = \bar{F}(w_1)/\mu_1 - 1/\mu_2 < 0$. Further, since $\xi_5 \geq 0$, we have $-\frac{\xi_5}{\lambda_2 H(w_2)} \leq 0$. Next, since $w_1 > w_2$ and $H(\cdot)$ is strictly decreasing, we have $\bar{F}(w_2)\left(\frac{1}{H(w_2)}-\frac{1}{H(w_1)}\right) < 0$ and $\bar{F}(w_1)\left(\frac{1}{H(w_2)}-\frac{1}{H(w_1)}\right) < 0$. Combining all these, we have $D - V_1\mu_1 - \frac{\mu_1}{H(w_1)} > 0$, thus contradicting the optimality condition for w_1. Hence, $w_1 > 0$ and $w_2 = 0$ cannot hold at the same time.

Consider next $w_1 = 0$ and $w_2 > 0$. Similar to the previous analysis, we have

$$D - V_2\mu_2 - \frac{\mu_2}{H(w_2)} = 0,$$
$$D - V_1\mu_1 - \frac{\mu_1}{H(w_1)} = -\xi_4 \frac{\mu_1}{\lambda_1 f(w_1)}.$$

where $\xi_4 \geq 0$ is the Lagragian multiplier corresponding to the constraint $w_1 \geq 0$ (see the proof of Proposition 2). Then,

$$D - V_2\mu_2 - \frac{\mu_2}{H(w_2)} = \frac{\frac{\xi_4}{\lambda_1 H(w_1)} + \frac{\bar{F}(w_2)}{H(w_2)} - \frac{\bar{F}(w_1)}{H(w_1)} + \int_{w_1}^{w_2} \bar{F}(y)dy}{\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2}. \tag{33}$$

Now,

$$\bar{F}(w_2)\left(\frac{1}{H(w_2)}-\frac{1}{H(w_1)}\right) < \frac{\bar{F}(w_2)}{H(w_2)} - \frac{\bar{F}(w_1)}{H(w_1)} + \int_{w_1}^{w_2} \bar{F}(y)dy < \bar{F}(w_1)\left(\frac{1}{H(w_2)}-\frac{1}{H(w_1)}\right).$$

Since $\frac{1}{H(w_2)} - \frac{1}{H(w_1)} > 0$ and $\xi_4 \geq 0$, the numerator of (33) is strictly positive. Irrespective of the sign of $\bar{F}(w_1)/\mu_1 - \bar{F}(w_2)/\mu_2$, $D - V_2\mu_2 - \frac{\mu_2}{H(w_2)}$ cannot be zero. Hence, $w_1 = 0$ and $w_2 > 0$ cannot hold at the same time. □

Proof of Lemma 5. Given μ_1 and μ_2, it is straightforward to verify that the equality in (20) corresponds to the first-order optimality condition for w, and the inequalities in (20) correspond to the constraints that $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. □

Proof of Corollary 3. By Lemma 2, we must have $w_1 = w_2 = w$ for an overloaded solution to be optimal. With $G(\mu_1, \mu_2, w, w) = \lambda_1 r(\mu_1, w) + \lambda_2 r(\mu_2, w)$, we have

$$\frac{\partial G}{\partial \mu_1} = \frac{\lambda_1 \bar{F}(w)}{\mu_1^2}\left(\frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2} + V'(\mu_1)\mu_1^2\right),$$
$$\frac{\partial G}{\partial \mu_2} = \frac{\lambda_2 \bar{F}(w)}{\mu_2^2}\left(\frac{V(\mu_1) - V(\mu_2)}{1/\mu_1 - 1/\mu_2} + V'(\mu_2)\mu_2^2\right),$$
$$\frac{\partial G}{\partial w} = \Lambda f(w)\left(-\frac{V(\mu_1)\mu_1 - V(\mu_2)\mu_2}{\mu_1 - \mu_2} - \frac{1}{H(w)}\right).$$

Following the same proof as of Proposition 4, it is easy to show that the optimal solution of $\mathcal{M}_2^{\mathrm{OL}}$ is achieved under $(\bar{\mu}, \underline{\mu})$ when $V'(\mu)\cdot\mu^2$ is strictly increasing. Noting the form of $\frac{\partial G}{\partial w}$ above, the result follows from the fact that the hazard rate of the patience distribution is decreasing. □

Proof of Proposition 8. The proof follows by considering our $\mathcal{S}$ policy for single-grade and two-grade systems separately and then combining them. Note first that the optimal grade allocation must follow the fluid solution. Otherwise, for any other allocation that deviates from the fluid solution, the resulting system welfare will deviate from the fluid solution by $\Omega(\Lambda)$ and it will be dominated by a policy that follows the grade allocation and scheduling policy in the fluid solution.

Next, for single-grade systems, if a critically-loaded regime is optimal (i.e., the optimal offered wait $w^* = 0$), then following Zeltyn and Mandelbaum (2005, Theorem 4.1), the expected queue length is within $\mathcal{O}(\sqrt{\Lambda})$ of the fluid approximation. If an overloaded regime is optimal (i.e., the optimal offered wait $w^* > 0$), then following Bassamboo and Randhawa (2010, Proposition 2), the expected queue length is within $o(1)$ of the fluid approximation. In both cases, because the patience distribution is exponential, the abandonment rate is within the same order of the fluid approximation as the queue length.

For two-grade systems, following Proposition 7, they must be critically loaded provided that they are optimal. Then, following the proof of Bassamboo et al. (2023, Proposition 6), both the abandonment rate and queue length are within $\mathcal{O}(\sqrt{\Lambda})$ of the fluid approximation for critically loaded systems.

Bassamboo and Randhawa (2010, Theorem 2) and Bassamboo et al. (2023, Proposition 6), respectively, establish that for single-grade and multi-grade systems, the fluid approximation gives a lower bound of the expected abandonment rate. Because the patience distribution is exponential, the fluid approximation also provides a lower bound for the expected queue length too. □

Proof of Lemma 6. For any feasible service-rate-first and waiting-time-next allocation $(\boldsymbol{\mu}, \boldsymbol{p}, \boldsymbol{w})$, we will construct a feasible waiting-time-first and service rate-next allocation $(\boldsymbol{w'}, \boldsymbol{p'}, \boldsymbol{\mu'})$ that achieves an identical objective value. The reverse follows analogously.

Consider an original allocation $(\boldsymbol{\mu}, \boldsymbol{p}, \boldsymbol{w})$, there are K customer grades and grade i has arrival rate Λp_i and service rate μ_i. This grade is further split into $J(i)$ subgrades and subgrade j has arrival rate $\lambda_{i,j}$ and offered wait $w_{i,j}$. Now, define

$$\boldsymbol{w'} := \left(w_{i,j} \middle| i \in \{1, 2, \ldots, K\}, j \in \{1, 2 \ldots, J(i)\} \right).$$

If $w_{i,j}$ is a singleton in this set, then define $w'_{i'} = w_{i,j}$ and $\Lambda p'_{i'} = \lambda_{i,j}$. Otherwise, suppose there exists $w_{i_1,j_1} = w_{i_2,j_2}$, then merge these two offered waits by setting $w'_{i'} = w_{i_1,j_1}$, $\Lambda p'_{i'} = \lambda_{i_1,j_1} + \lambda_{i_2,j_2}$, $\mu_{i',1} = \mu_{i_1}$, $\mu_{i',2} = \mu_{i_2}$, and service rate allocation probability $\lambda_{i_1,j_1}/(\lambda_{i_1,j_1} + \lambda_{i_2,j_2})$ and $\lambda_{i_2,j_2}/(\lambda_{i_1,j_1} + \lambda_{i_2,j_2})$, respectively. The new allocation $(\boldsymbol{w'}, \boldsymbol{p'}, \boldsymbol{\mu'})$ is a feasible solution to the waiting-time-first and service rate-next allocation problem because it needs the same total capacity as the original solution. Further, because each $\lambda_{i,j} r(\mu_i, w_{i,j})$ term in the original objective function has an equivalent counterpart $\lambda_{i',j'} r(\mu_{i',j'}, w_{i'})$ in the new objective function, the new allocation $(\boldsymbol{w'}, \boldsymbol{p'}, \boldsymbol{\mu'})$ achieves the same system welfare as the original solution. □

Proof of Proposition 9. Assume $\Lambda_i > 0$ for all classes with arrivals; classes with $\Lambda_i = 0$ can be omitted. First, eliminate the grade-allocation probabilities and capacity variables. The two-stage

problem (21)–(22) is equivalent to

$$\begin{aligned} \sup_{\{\mu_{i,a},w_{i,a},\lambda_{i,a}\}} \quad & \sum_{i=1}^{m}\sum_{a\in A_i} \lambda_{i,a} r_i(\mu_{i,a}, w_{i,a}) \\ \text{s.t.} \quad & \sum_{a\in A_i} \lambda_{i,a} = \Lambda_i, \qquad i = 1,\dots,m, \\ & \sum_{i=1}^{m}\sum_{a\in A_i} \lambda_{i,a}\frac{\bar{F}_i(w_{i,a})}{\mu_{i,a}} \le N, \\ & \lambda_{i,a} \ge 0. \end{aligned} \tag{34}$$

Here a indexes the service rate and offered wait pairs created within class i. The equivalence follows as in Proposition 1: any feasible solution of the original formulation induces (34) by setting $a = (j,k)$, $\mu_{i,a} = \mu_{i,j}$, $w_{i,a} = w_{i,j,k}$, and $\lambda_{i,a} = \lambda_{i,j,k}$. Conversely, any feasible solution of (34) can be implemented in the original formulation by assigning class-i arrivals to pair a with probability $\lambda_{i,a}/\Lambda_i$ and allocating capacity $n_{i,a} = \lambda_{i,a}\bar{F}_i(w_{i,a})/\mu_{i,a}$. This preserves feasibility and objective value.

Now fix an optimal solution of (34), and condition on its finite set of candidate pairs $\{(\mu^*_{i,a}, w^*_{i,a})\}$. Define $v_{i,a} := r_i(\mu^*_{i,a}, w^*_{i,a})$ and $q_{i,a} := \frac{\bar{F}_i(w^*_{i,a})}{\mu^*_{i,a}}$. Given these pairs, the remaining optimization over arrival rates is the linear program

$$\begin{aligned} \max_{\{\lambda_{i,a}\ge 0\}} \quad & \sum_{i=1}^{m}\sum_{a\in A_i} \lambda_{i,a} v_{i,a} \\ \text{s.t.} \quad & \sum_{a\in A_i} \lambda_{i,a} = \Lambda_i, \qquad i = 1,\dots,m, \\ & \sum_{i=1}^{m}\sum_{a\in A_i} \lambda_{i,a} q_{i,a} \le N. \end{aligned} \tag{35}$$

The fixed optimal solution is feasible and optimal for (35); otherwise, an improved solution to (35) would improve the original fluid problem.

By standard linear programming theory, (35) has an optimal basic feasible solution. The constraint matrix has m class-balance constraints and one aggregate capacity constraint, so its rank is at most $m+1$. Hence an optimal basic feasible solution has at most $m+1$ positive arrival-rate components $\lambda_{i,a}$. Since $\sum_{a\in A_i}\lambda_{i,a} = \Lambda_i > 0$, each class must have at least one positive component. Therefore, across m classes, at most one class can have more than one positive component; moreover, no class can have more than two positive components. Thus, at most one class uses two active service rate and offered wait pairs, and every other class uses one.

Mapping this basic optimal solution back to (21)–(22) gives the desired optimal grade-allocation and scheduling solution. □

Proof of Proposition 10. The proof follows the same structure as the proof of Proposition 8. The fluid upper bound $K^*_\Lambda \le \mathcal{M}^*(N,\Lambda)$ is independent of whether the fluid solution is implemented through the $\mathcal{S}$ or $\mathcal{W}$ policy, because the fluid problem is a relaxation of the stochastic control

problem. Also, any grade allocation that deviates from the fluid solution loses $\Omega(\Lambda)$ welfare relative to a policy that follows the fluid allocation and scheduling rule.

It remains to show that the $\mathcal{W}$ policy achieves the fluid value up to $\mathcal{O}(\sqrt{\Lambda})$.

Consider first the single-grade case. In this case, service rate differentiation is irrelevant, so the $\mathcal{W}$ policy coincides with the $\mathcal{S}$ policy. Hence the same Markovian many-server approximation bounds used in Proposition 8 apply directly.

Next consider the two-grade case. Following Proposition 7, a two-grade fluid solution can be optimal only if it is critically loaded, with $w_1^* = w_2^* = 0$. Let the optimal fluid solution be $(\mu_1^*, \mu_2^*, \lambda_1^*, \lambda_2^*)$, and set $p_i^* = \lambda_i^*/\Lambda$ for $i = 1, 2$. Because $N = \Lambda/\mu_0$, the normalized critically loaded two-grade fluid problem is independent of Λ, so we fix an optimal mixture $(p_1^*, p_2^*, \mu_1^*, \mu_2^*)$ along the sequence.

Under the two-grade $\mathcal{W}$ policy, a customer selected for service is assigned service rate μ_i^* with probability p_i^*. Hence the service time distribution is a two-phase hyperexponential distribution, equivalently a phase-type distribution with initial distribution (p_1^*, p_2^*) and transient rates (μ_1^*, μ_2^*). Its mean is

$$\sum_{i=1}^{2} \frac{p_i^*}{\mu_i^*} = \frac{1}{\Lambda}\left(\frac{\lambda_1^*}{\mu_1^*} + \frac{\lambda_2^*}{\mu_2^*}\right) = \frac{N}{\Lambda} = \frac{1}{\mu_0}.$$

Thus the induced many-server system is critically loaded at the fluid scale.

Dai et al. (2014) state their result for a FIFO $GI/Ph/n + M$ queue, and the aggregate state process under the critically loaded $\mathcal{W}$ policy has the same law as that FIFO system. Indeed, under exponential patience, the abandonment rate is γQ when Q customers are waiting, independently of their order in queue; and, at each service entry, the initial service phase is sampled independently according to (p_1^*, p_2^*). Therefore the transition rates of the aggregate process consisting of the queue length and the service phase counts do not depend on whether the waiting customer is selected by LCFS or FCFS.

We can therefore apply the steady-state Halfin–Whitt bound for $GI/Ph/n + M$ queues in Dai et al. (2014, Theorem 1 and the argument following Lemma 4). Let $Q_\Lambda^{\mathcal{W}}$ denote the steady-state queue length under the $\mathcal{W}$ policy. Their Lyapunov bound implies that the diffusion-scaled steady-state queue length is uniformly integrable; hence, $\mathbb{E}[Q_\Lambda^{\mathcal{W}}] \le B\sqrt{N} = O(\sqrt{\Lambda})$ for some finite constant B. Because patience times are exponential, the steady-state abandonment rate is $\gamma\mathbb{E}[Q_\Lambda^{\mathcal{W}}]$, and the total service-entry rate is $\Lambda - \gamma\mathbb{E}[Q_\Lambda^{\mathcal{W}}]$. Since service rates are assigned independently at service entry, the rate at which service rate μ_i^* is used is $p_i^*\left(\Lambda - \gamma\mathbb{E}[Q_\Lambda^{\mathcal{W}}]\right)$ for $i = 1, 2$.

Let $V_i = V(\mu_i^*)$. The two-grade critically loaded fluid value is $\mathcal{M}_2^{\mathrm{CL}} = \Lambda \sum_{i=1}^2 p_i^* V_i$, whereas the stochastic welfare under the $\mathcal{W}$ policy is $K_\Lambda^{\mathcal{W}} = \left(\Lambda - \gamma\mathbb{E}[Q_\Lambda^{\mathcal{W}}]\right)\sum_{i=1}^2 p_i^* V_i - c\,\mathbb{E}[Q_\Lambda^{\mathcal{W}}]$. Therefore,

$$\left|\mathcal{M}_2^{\mathrm{CL}} - K_\Lambda^{\mathcal{W}}\right| \le \left(\gamma \max_{i=1,2}|V_i| + c\right)\mathbb{E}[Q_\Lambda^{\mathcal{W}}] = O(\sqrt{\Lambda}),$$

because $V(\cdot)$ is bounded on the compact feasible service rate set.

Combining the single-grade and two-grade cases, there exists a finite constant $A > 0$ such that

$\mathcal{M}^*(N,\Lambda) - K_\Lambda^{\mathcal{W}} \le A\sqrt{\Lambda}$. Since $K_\Lambda^{\mathcal{W}} \le K_\Lambda^* \le \mathcal{M}^*(N,\Lambda)$, it follows that

$$0 \le K_\Lambda^* - K_\Lambda^{\mathcal{W}} \le A\sqrt{\Lambda}.$$

Finally, if $\mathcal{M}^*(N,\Lambda)$ is achieved under (11) with $w^* > 0$, then the optimal fluid solution is single-grade and overloaded. In this case the $\mathcal{W}$ policy coincides with the $\mathcal{S}$ policy, and Bassamboo and Randhawa (2010, Proposition 2) implies $\mathcal{M}^*(N,\Lambda) - K_\Lambda^{\mathcal{W}} = o(1)$. Together with $K_\Lambda^{\mathcal{W}} \le K_\Lambda^* \le \mathcal{M}^*(N,\Lambda)$, this gives $\left|K_\Lambda^* - K_\Lambda^{\mathcal{W}}\right| \to 0$.

□