# The principal-agent problem for service rate event-dependency

Benjamin Legros

# The principal-agent problem for service rate event-dependency

Benjamin Legros

Ecole de Management de Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France

benjamin.legros@centraliens.net

**Abstract**

This study aims to determine the cost of letting an agent adjust the service rate to the last realized event, being a customer arrival or a service completion. We study this question in a single-server queue under a principal-agent framework. The principal seeks to reduce the expected waiting time by incentivizing the agent to modify the service rate through a performance-based payout. We show that a large range of improvement is achievable by selecting event-dependent service rates. However, the agent's payout can grow high in the realized improvement, suggesting to limit the use of incentives for event-dependent service rates to a bounded waiting time improvement. When the service rate after an arrival is contractible, the agent should be paid more in contexts with a low variability inter-arrival time. The opposite conclusion holds when the average service rate is contractible. Further, we provide a criterion to determine when it is optimal for the agent to accelerate after an arrival or after a service completion. Finally, we investigate the effect of event-dependency on customers' fairness and abandonment.

**Keywords:** Queueing; event-dependency; G/M/1; service rate; principal-agent.

## 1 Introduction

In some manufacturing and service systems, agents are able to decide on their service rate. In queue-modeled systems, for instance, an agent can optimize its service rate to obtain the best trade-off between the idling duration and effort-cost. Depending on the system under consideration, the optimized service rate is either constant, or it can be changed over time depending on the queue length (George and Harrison, 2001) or the objective service level (Zafer and Modiano, 2005). In studies related to service rate optimization, the rational and strategic agent is assumed to have a full understanding of the system state and is subsequently able to adjust the service rate to a certain optimization objective.

In practice, a human agent may be subject to some bias in perceiving the situation of the system. One of these biases is to give particular importance to the most recent changes in the system, more than to the system state itself. This bias is regularly experienced by newspaper readers, for whom

the latest news tends to preempt older news in their understanding of a situation. In Finance, decisions are also often made with a myopic bias (Fellner and Sutter, 2009). In a queueing context, the last realized event may influence the agent in selecting the service speed. When an arrival occurs, the agent might be tempted to accelerate to quickly get back to the pre-arrival situation, for instance. A service completion could also be an incentive to speed up as it indicates that the idling period could then come sooner than expected. Delasay et al. (2019) reviewed how changeover, instantaneous load, and extended load may impact service times. The effect of the last realized event can be viewed as a reaction to a modification of the load (extension or reduction) on the service rate.

Capturing the effect of the last realized event from data analysis can be confusing as the last event may not necessarily be the cause of the agent's behavior. For instance, the service duration of customers who did not obtain service completion (from another agent) during their service could be evaluated. Having no service completion over a long period may indicate that only a small number of agents is busy. This means that busy agents are aware that they are likely to become available at service completion and could thus accelerate their service speed. In this case, what drives the decision is not that there is no service completion during the service time but that some other agents could be available. On the other hand, a high frequency of service completion could indicate that the queue is congested. Therefore, there exists a correlation between the last realized event and the system state. This means that the proven rationality when taking decisions based on queue length may also exist when taking decisions based on the last realized event. The aim of this paper is to analyze the costs and benefits of taking decisions for the service speed driven by the last realized event, being an arrival or a service completion.

To this end, we investigate a principal-agent problem where the agent is hired to serve customers arriving over time from a process with a generally distributed inter-arrival time. The service times are exponentially distributed and the agent can select two different service rates depending on the last realized event. A fixed payment of the agent is due to a base-effort exerted to serve all customers at a given base service rate. However, the principal expects the agent to speed up service as compared to this base-effort with the aim of reducing customers' expected waiting time. To achieve this goal, the principal uses incentives to reward the improvement in expected waiting time. The agent reacts by selecting event-dependent service rates, that are different than those of the

base-effort, to accelerate some services. Two contracts are studied. In this first one, the service rate after an arrival is contractible. Thus, the incentive only has an effect on the service rate after a service completion which could be increased by the agent to lower the waiting time. In the second contract, the average service rate is contractible and the agent has full discretion to select the two service rates. For the two contracts, the agent selects the service rates to maximize the utility function defined as the difference between the revenue and effort-cost.

To solve the principal-agent problem, we first determine the stationary probabilities and the performance measures of the G/M/1 queue with event-dependent service rates. To this end, we employ a continuous time Markov chain analysis similar to the one in Kerner (2008) for an M/G/1 queue, where a state of the system is defined by the number of customers present, the nature of the last realized event, and the remaining time before the next arrival. Next, for the first contract, we prove that the agent's utility is concave in the obedient effort, being the average service rate or the service rate after a service completion. This result indicates that the agent's utility has a unique maximum which is solution of the optimization problem. Further, we show how the payout parameters should be set to minimize the agent's payout while achieving a certain level of improvement on the expected waiting time. Although the range of achievable improvement is large with the first contract, our results show that the agent's payout is convex in the expected waiting time improvement. This suggests that improving the service level through incentives should be limited to relatively moderate expectations in the waiting time lowering. Finally, we show that the agent's payout reduces with the variability of the inter-arrival time, indicating that selecting event-dependent service rates may be a good strategy to reduce the negative consequences of inter-arrival time variability on the expected waiting time.

For the second contract, where the average service rate is contractible, we provide a condition on the inter-arrival time distribution under which it is optimal for the agent to accelerate after an arrival or service completion. This condition is complex and involves the Laplace Transform of the inter-arrival time distribution. However, in many cases, when the variability of the inter-arrival time distribution increases, it is optimal for the agent to accelerate after an arrival. Moreover, we also prove that having an infinite service rate after an arrival or service completion, in addition to not being feasible in practice, is also never optimal. Using these results, we express the agent's payout parameters. With the second contract, the range of achievable improvement is less than

with the first contract as the expected waiting time is mainly driven by the average service rate. The behavior of the agent's payout in the waiting time improvement differs from the one in the first contract. When the variability of the inter-arrival time is high, the payout can be concave when the improvement is low. This argues for not employing event-dependent service rates in this case. On the contrary with low variability distributions, the agent's payout is almost insensitive to the waiting time improvement which renders event-dependent service rates advisable in such cases.

Finally, we investigate other benefits of event-dependent service rates in the case of an exponential inter-arrival time distribution. We show that accelerating after a service completion can reduce the sensitivity of the expected waiting time at arrival to queue length. This can be interesting when the principal aims to provide a fair service quality to its customers in the sense that arriving customers have almost the same offered waiting time at arrival. We also include the feature of abandonment in the analysis. We show that accelerating after an arrival can help to reduce the fraction of lost customers.

Section 2 presents a literature review. Section 3 formulates the principal-agent optimization problem for service rate event-dependency. Section 4 derives the stationary probabilities and the performance measures of the G/M/1 queue with event-dependent service rates. Section 5 provides the solution of the optimization problem when the service rate after an arrival is contractible while Section 6 explores the solution to this problem when the average service rate is contractible. Section 7 reveals other benefits of event-dependency on the waiting time at arrival and fraction of lost customers in the exponential case. Finally, Section 8 concludes the paper and provides directions for future research. The mathematical proofs of the main results are given in the appendix.

## 2 Literature review

First, since our study considers a principal-agent problem, we present prior studies in the related field of contract theory. Second, in relation with service rate optimization, we provide the existing literature on agents acting strategically. Finally, as our analysis examines a particular G/M/1 queue, we detail the literature on methodological contributions on this topic.

The book by Laffont and Martimort (2009) provides a theoretical framework to analyze the principal-agent problem considered in this study. There is a long history of compensation analyses in the fields of marketing, economics, heath care, and operations management (for instance, see Lal

and Srinivasan (1993); Herweg et al. (2010); Jain (2012); Chen et al. (2016); Suen et al. (2018); Li et al. (2020)). This literature stream mainly focuses on linear commission and quota-bonus contracts (that is, the employee receives a bonus for meeting a performance quota). In this paper, we focus on a performance-based payout used as a tool to induce a lowering of the offered wait. We assume that either one of the two service rates is contractible or the average service rate is contractible. Suen et al. (2018) describes how payouts can be employed to induce socially-optimal behavior in a context where decisions are non-contractible, as in our study. Jain (2012) investigated a situation where employees exert self-control, as in this paper, showing that firms can reduce the negative consequences of self-control by delaying payment to employees. Jiang et al. (2012) investigated a performance-based approach to contracting for medical services. They showed that including the waiting time in the contract definition through a penalty paid in case of excessive wait allows reaching the first-best performance. The contract considered in this study is also a performance based contract. However, instead of a penalty for excessive wait, we consider a reward for wait reduction effort. In a producer-seller relationship, Chen et al. (2016) compared forecast-based and linear compensation contracts. They showed that with an endogenous information-acquisition effort, forecast-based contracts can outperform linear compensation ones. Li et al. (2020) also compared linear and non-linear contracts and showed that the feature of fairness plays a role in the potential outcomes realized, leading to a reduction in the benefits of non-linear contracts. Meanwhile, Long and Nasiry (2020) discussed contexts where making wages transparent to employees were beneficial to the firms. The incentive-design issue becomes more complicated with multitasking agents. Dai et al. (2021) considered a principal-agent framework where the agent can exercise two types of tasks, operational and marketing. They characterized the optimal compensation plan, where a bonus is paid when either all the inventory above a threshold is sold or the sales quantity meets an inventory-dependent target.

This paper analyzes the problem of an agent acting strategically to select its service speed, tying our study to queueing games (Hassin and Haviv, 2003). It should be noted that service rate optimization in the literature can also be made by the system manager (Ata and Shneorson, 2006). Kalai et al. (1992) investigated a queue with two exponential servers in competition. They proved that when the expected waiting time is finite, there exists a unique symmetric strategic equilibrium. Christ and Avi-Itzhak (2002) extended the model description to a situation with a state-dependent

Poisson arrival process, showing that when the cost function is convex and increasing, there also exists a unique symmetric Nash equilibrium strategy. Avi-Itzhak et al. (2006) explored globally optimal solutions for this model. They showed that optimal solutions are symmetric and that the unique Nash equilibrium is in general strictly inferior to a globally optimal solution. Cachon and Zhang (2007) studied a two-server model where the service rate selection is driven by incentives as in this study. They showed that the trade-off between efficiency and incentives may not exist. In particular, it seems possible to design an allocation policy that is efficient and leads the faster-working servers. Geng et al. (2015) focused on the impact of exogenous routing decisions on endogenous service speed in a queue with two servers seeking fairness. The existence and uniqueness of the Nash equilibrium was proven for some routing policies. Zhan and Ward (2018) analyzed a single-server exponential queue with abandonment. They proved that there exists a unique maximum of the agent's utility computed as the product of the value of the service speed multiplied by the agent's utilization rate. In the multi-server case, Gopalakrishnan et al. (2016) considered a situation where each agent can decide on their service rate, given that the agent's objective is to maximize utility defined as the difference between the fraction of idling time and the effort-cost. The authors also discussed the effect of routing rules on the optimal service rate. Chan et al. (2014) studied how to optimally speed up the agents' service rate to reduce congestion in a system where customers with a too short service time may return. Zhan and Ward (2019) further investigated the multi-server case to find a joint staffing, routing, and payment policy that would lead to an optimal performance. By solving the centralized control problem under fluid scaling, they found that critically loaded, efficiency driven, quality driven, and intentional idling regimes were economically optimal. Finally, in a principal-agent framework close to ours, Baiman et al. (2010) investigated how a single server could maximize its utility in a finite capacity queue by selecting a service rate. The principal in their context could decide for the payouts and for the system's capacity.

This paper derives the stationary probabilities in a particular G/M/1 queue. The G/M/1 queue is one of the classical models of queueing theory (for instance, see Kleinrock (1975), Chapter 6). The usual approach to derive the performance measures for this queue is to analyze the related discrete time Markov chain at arrival instants. This method has been successful in providing the performance measures for some variants of the G/M/1 queue. Laslett (1975) used this approach to analyze the G/M/1 with finite capacity, while Hokstad (1975) extended the results to the multi-server setting.

In addition, Zhang and Tian (2004) considered a queueing system in which the server follows a threshold-type policy. In such a system, the server stops serving the queue whenever the system becomes empty and resumes service when the number of waiting customers in the system reaches a certain threshold. Alternative approaches like martingale techniques, transform techniques, and sample-path arguments are developed in Adan et al. (2005) to analyze the G/M/1 queue. Pourbabai (1990) developed a heuristic algorithmic approach to approximate the tandem behavior of the finite G/M/1 queue, while Ke and Wang (2002) used the supplementary variable technique where the remaining inter-arrival time becomes the supplementary variable to analyze the G/M/1 queue with a removable server. We use this approach in our paper to compute the stationary probabilities. Chae et al. (2006) derived the probability-generating function of the queue length when the server takes one exponential vacation each time the system empties. Legros (2021) developed a model where the idling duration and the age of the oldest customer is used to provide a Markovian representation of the G/M/1 queue with the aim of solving policy optimization problems. Bae and Kim (2010) employed level crossing arguments to analyze the G/M/1 queue with constant patience. Also using a sample path analysis and level crossing arguments, Löpker and Perry (2010) investigated the idle periods of the G/M/1 queue with a removable server. Haviv and Kerner (2011) showed that for the G/M/1 queue, conditioning on a busy server, the age of the inter-arrival time and the number of customers in the queue are independent and that the same result holds when the age is replaced by the residual inter-arrival time. Oz et al. (2017) introduced a rate balance principle for general stochastic processes and used this result to derive new results for G/M$_n$/1 queueing systems.

## 3 Formulation of the problem

In this section, we present the principal-agent framework of this study. The firm consists of a risk-neutral principal and agent who agree to a contract which will govern their employment relation. The agent is hired to serve customers in the order of their arrival in the system. Customers' inter-arrival time is exogenous and generally distributed with probability density function $f(t)$, for $t \geq 0$. Further, we assume that the service time is distributed according to an exponential distribution. We assume that the agent does not know the total number of customers waiting to be served and cannot anticipate future arrivals. However, a signal is sent by the system to the agent when a new customer enters the system. Therefore, the agent knows customers' arrival instants. Since the

agent is the only one to serve customers, service completion instants are also known by the agent. Consequently, even if this information is not communicated, the agent could determine the number of customers in the system from the arrival and service completion instants. However, we assume that the agent does not make this effort and only relies on the last realized event being either an arrival or a service completion to select a service rate. Therefore, the agent may select two different service rates, $\mu_1$ and $\mu_2$, such that the rate $\mu_1$ is used after an arrival instant and $\mu_2$ is used after a service completion. It should be noted that the service rate can then change during a service if an arrival occurs. This queueing model is a novel variant of the G/M/1 queue, termed the $G/M^{\text{event}}/1$ queue, where the service rate is event-dependent. We denote by $\mu$ the average service rate of a given customer and by $E(A)$ the expected inter-arrival time. As for the G/M/1 queue, the stability condition is given by $\mu E(A) > 1$.

The compensation model is based on a fixed-wage $FW$ and a piece-rate compensation $pr$ based on the long-run performance offered to customers. The fixed-wage compensates the base-effort of working with an average service rate $\mu$ while the piece-rate compensation rewards the extra-effort which results in reducing the expected waiting time, $E(W)$, by selecting different service rates, $\mu_1$ and $\mu_2$. The principal rewards the improvement in the expected waiting time when the agent selects different service rates than those of a base situation where the agent would operate with identical service rates, $\mu_1 = \mu_2 = \mu_e$. Specifically, the agent receives a reward $pr$ per customer and per unit of expected waiting time gained as compared to the expected waiting time in the base situation. In the long-run, the expected revenue for the agent, termed $C$, is then $C = FW + \frac{pr}{E(A)}(E(W_e) - E(W))$, where $E(W_e)$ is the expected waiting time in the base situation (that is, when $\mu_1 = \mu_2 = \mu_e$). The agent's effort is the sum of a base-effort, $b\mu$, of working with an average service rate $\mu$ and an extra-effort, $e\,(\max(\mu_1, \mu_2) - \mu)$, of selecting different service rates, with $b, e \geq 0$. Further, we assume that the extra-effort is more costly to the agent than the base-effort: $e \geq b$. The agent's utility, $U$, can then be defined as the difference between the expected revenue and effort-cost: $U = FW + \frac{pr}{E(A)}(E(W_e) - E(W)) - e\max(\mu_1, \mu_2) + (e - b)\mu$. The agent selects $\mu_1$ and $\mu_2$ such that the long-run expected utility is maximized. The principal has discretion for selecting the compensation parameters $pr$ and $FW$. However, the fixed-wage $FW$ should compensate the base-effort. Therefore $FW$ is set such that $FW \geq b\mu$, for $b \geq 0$.

It should be noted that the most standard contract definition for principal-agent problems is to

reward the service of each customer as in Baiman et al. (2010). In their case, the agent is incentivized to speed-up service as a way to serve more customers since they consider a finite capacity queue. In our study, the queue capacity is infinite, so the selection of the service rate does not impact the long-run rate of served customers. That is why the service level, measured by the waiting time, is included in the contract definition as a way to incentivize the agent to speed up. This choice can be found in different contexts. We mention Ren and Zhou (2008) who considered a contract between a user company and a call center. They proved that quantity based contracts are not optimal while contracts involving the service level could be. In the health care sector, the wait is also included in the contract structure. For instance, Jiang et al. (2012) investigated a performance-based approach to contracting for medical services. They showed that including the waiting time in the contract definition through a penalty paid in case of excessive wait allows reaching the first-best performance. In this study, instead of a penalty for excessive wait, we consider a reward for wait reduction effort. Finally, Guo et al. (2019) included the waiting time in their utility function which is also part of the reward structure in their considered contracts.

The principal-agent optimization problem can be expressed as follows:

$$\begin{aligned} &\underset{pr,FW}{\text{Minimize}}\ FW + \frac{pr}{E(A)}(E(W_e) - E(W)), \\ &\text{subject to} \begin{cases} FW \geq b\mu \\ \mu_1, \mu_2 \in \arg\max\{FW + \frac{pr}{E(A)}(E(W_e) - E(W)) - e\max(\mu_1, \mu_2) + (e-b)\mu\}, \end{cases} \end{aligned} \tag{1}$$

We consider two contracts to analyze Problem 1. In the first one, termed Contract 1, we assume that the rate $\mu_1$ is kept constant such that Problem 1 is solved by determining the value of $\mu_2$ which maximizes the agent's utility. Under this constraint, the agent's actions are partially contractible. Specifically, the rate $\mu_1$ is contractible while $\mu_2$ is not. The objective for the principal with this contract is to minimize the agent's payout such that a certain improvement in $E(W)$ is reached by increasing $\mu_2$. It should be noted that the analysis with Contract 1 could be made in a similar way by assuming that $\mu_2$ would be kept constant while $\mu_1$ would be optimized. With the second contract, termed Contract 2, we assume that the average service rate $\mu$ is contractible but the two service rates $\mu_1$ and $\mu_2$ can be optimized. This second contract focuses more specifically on the effect of the differentiation between the service rates as the average service rate is kept constant. As for Contract 1, the objective for the principal is to achieve a certain improvement on $E(W)$ at

minimal cost. To solve Problem 1, we first evaluate the performance measures of the $G/M^{\text{event}}/1$ queue. This analysis is made in Section 4 using the supplementary variable approach as in Kerner (2008). Next, in Sections 5 and 6, we solve Problem 1 for the two contracts described above. Finally, in Section 7, we explore the impact of having different service rates on the expected waiting time at arrival and proportion of abandonment. We end this section with a table of notations used throughout the paper (Table 1).

Table 1: Table of notations

| | |
|---|---|
| **State of the system** | |
| $x$ | Number of customers in the system, $x \in \mathbb{Z}^+$ |
| $i$ | Last realized event with $i = 1$ for an arrival and $i = 2$ for a service completion |
| $r$ | Remaining inter-arrival time, $r \in \mathbb{R}^+$ |
| **Parameters of the queueing model** | |
| $f(t)$ | Probability-density function of the inter-arrival time |
| $F^*(s)$ | Laplace-Stieltjes transform of $f(t)$: $F^*(s) = \int_{r=0}^{\infty} e^{-sr} f(r)\,\mathrm{d}r$ |
| $E(A)$ | Expected inter-arrival time |
| $\mu_1$ | Service rate after an arrival |
| $\mu_1$ | Service rate after a service completion |
| $\mu$ | Average service rate |
| $\mu_e$ | Service rate in the base situation with $\mu_e = \mu_1 = \mu_2$ |
| **Agent's utility, payments and effort** | |
| $pr$ | Piece-rate compensation per customer and per unit of expected waiting time gained |
| $FW$ | Fixed-wage per agent |
| $b$ | Parameter of the base-effort (Base-effort= $b\mu$) |
| $e$ | Parameter of the extra-effort (Extra-effort= $e(\max(\mu_1, \mu_2) - \mu)$) |
| $U$ | Agent's utility, $U = FW + \frac{pr}{E(A)}(E(W_e) - E(W)) - e\max(\mu_1, \mu_2) + (e - b)\mu$ |
| $C$ | Agent's revenue (that is, the principal's cost), $C = FW + \frac{pr}{E(A)}(E(W_e) - E(W))$ |
| **Performance measures** | |
| $p(x, i, r)$ | Probability density to be in state $(x, i, r)$ |
| $\pi_{x,i}$ | Stationary probability to be in state $(x, i)$ |
| $q$ | Common ratio of the stationary probabilities $\pi_{x,i}$ |
| $E(W)$ | Expected waiting time |
| $E(W_e)$ | Expected waiting time in the base situation with $\mu_e = \mu_1 = \mu_2$ |
| $q_e$ | Common ratio of the stationary probabilities $\pi_{x,i}$ in the base situation with $\mu_e = \mu_1 = \mu_2$ |
| $H_x^i$ | Laplace transform of the first passage time from state $(x, i)$ to a beginning of service |
| $E(W_x)$ | Expected waiting time at arrival when $x$ customers are already present in the system |
| $\beta$ | Abandonment rate |
| $P_A$ | Probability of abandonment |

# 4 Performance analysis of the $G/M^{\text{event}}/1$ queue

The traditional framework to analyze queues with generally distributed inter-arrival times and exponential service times is to consider the embedded Markov chain at arrival instants (for instance, see Kleinrock (1975), Chapter 6, page 241). The analysis of the related discrete time Markov chain

may allow us to determine the stationary probabilities at arrival instants. For the G/M/1 queue, it is possible to deduce some important performance metrics like the expected waiting time or probability of having an empty system. In our case involving event-dependent service rates, however, the service time distribution may change during a service. Therefore, the expected waiting time cannot be deduced from the stationary probabilities at arrival instants. The same difficulty holds when considering the probability of having an empty system as the system at arbitrary time cannot be easily related to the system at arrival instants. To avoid this difficulty, we directly study the continuous time Markov chain at arbitrary instants by including the remaining time before the next arrival in the state description. This alternative approach was shown to be successful for specific M/G/1 and G/M/1 queues where the analysis of the embedded Markov chain does not lead to the wanted performance measures (Ke and Wang, 2002; Kerner, 2008; Legros and Sezer, 2018).

At a given instant $t$, a state of the system is defined by the vector $(x(t), i(t), r(t))$, where $x(t)$ is the number of customers in the system, with $x(t) \in \mathbb{Z}^+$, $i(t)$ is the last realized event, with $i(t) \in \{1, 2\}$, where $i(t) = 1$ ($i(t) = 2$) indicates that the last event is an arrival (a service completion), and $r(t)$ is the remaining time before the next arrival, with $r(t) \in \mathbb{R}^+$. At an arbitrary instant, a single service completion occurs within $\delta t$ time units with probability $\mu_1 \delta t + o(\delta t)$ or $\mu_2 \delta t + o(\delta t)$, two or more service completions occur with probability $o(\delta t)$ and the probability of no service completion is either $1 - \mu_1 \delta t + o(\delta t)$ or $1 - \mu_2 \delta t + o(\delta t)$. The vector $(x(t), i(t), r(t))$ is therefore a Markov process since it completely summarizes all past history relevant to the future system development.

We denote by $p_t(x, i, r)$ the probability-density of having $x$ customers in the system, $x \geq 0$, the last event being $i$, for $i = 1, 2$, and a remaining time before the next arrival being $r$, for $r \geq 0$, at time $t$ (given some arbitrary initial distribution), and by $p(x, i, r)$ the limit of $p_t(x, i, r)$ as $t$ tends to infinity; $p(x, i, r) = \lim_{t \longrightarrow \infty} p_t(x, i, r)$, for $x \geq 0$, $r \geq 0$, and $i = 1, 2$. In Lemma 1, we give the differential equations defining the evolution of the system state when the stationary regime is reached.

**Lemma 1.** *For $r \geq 0$, $x \geq 0$, and $i = 1, 2$, $p(x, i, r)$ obeys the following differential equations*

$$p(0,2,r)' = -\mu_2 p(1,2,r) - \mu_1 p(1,1,r), \tag{2}$$

$$p(x,2,r)' = \mu_2 p(x,2,r) - \mu_2 p(x+1,2,r) - \mu_1 p(x+1,1,r), \text{ for } x \geq 1, \tag{3}$$

$$p(1,1,r)' = \mu_1 p(1,1,r) - f(r)p(0,2,0), \text{ and} \tag{4}$$

$$p(x,1,r)' = \mu_1 p(x,1,r) - f(r)(p(x-1,1,0) + p(x-1,2,0)), \text{ for } x \geq 2, \tag{5}$$

*where $p(x, i, r)' = \frac{\partial p(x,i,r)}{\partial r}$.*

In Theorem 1, we obtain the steady state probabilities $\pi_{x,i}$ from Lemma 1, where $x$ represents the number of customers in the system and $i$ is the last realized event, where $i = 1$ ($i = 2$) indicates that the last realized event is an arrival (a service completion). Note that if the last realized event is an arrival, we cannot have an empty system (that is, we cannot have the combination $i = 1$ with $x = 0$).

**Theorem 1.** *Under the stability condition $\frac{\mu_1 F^*(\mu_1) + \mu_2(1 - F^*(\mu_1))}{\mu_1 \mu_2} < E(A)$, the steady state probabilities are given by*

$$\pi_{0,2} = 1 - \frac{\mu_1 F^*(\mu_1) + \mu_2(1 - F^*(\mu_1))}{\mu_1 \mu_2 E(A)}, \tag{6}$$

$$\pi_{x,2} = \frac{F^*(\mu_1)\mu_1(1-q)}{\mu_2(\mu_1 E(A) - 1) + (\mu_2 - \mu_1)F^*(\mu_1)} q^{x-1} \pi_{0,2}, \text{ for } x \geq 1, \tag{7}$$

$$\pi_{x,1} = \frac{(1 - F^*(\mu_1))\mu_2(1-q)}{\mu_2(\mu_1 E(A) - 1) + (\mu_2 - \mu_1)F^*(\mu_1)} q^{x-1} \pi_{0,2}, \text{ for } x \geq 1, \tag{8}$$

*where $F^*(s)$ the Laplace-Stieltjes Transform (LST) of the inter-arrival time; $F^*(s) = \int_{r=0}^{\infty} e^{-sr} f(r)\, \mathrm{d}r$, and $q$ is the unique solution in $(0, 1)$ of*

$$\frac{(\mu_1 - \mu_2)(1-q)F^*(\mu_1) + q\mu_1 F^*(\mu_2(1-q))}{\mu_1 - \mu_2(1-q)} = q. \tag{9}$$

From the stationary probabilities, we deduce the performance measures of interest in Corollary 1.

**Corollary 1.** *The main performance measures are given by*

$$\mu = \frac{\mu_1\mu_2}{\mu_1 F^*(\mu_1) + \mu_2(1 - F^*(\mu_1))}, \tag{10}$$

$$\pi_{0,2} = 1 - \frac{1}{\mu E(A)}, \text{ and} \tag{11}$$

$$E(W) = \frac{q}{\mu(1-q)}. \tag{12}$$

**The exponential case.** We illustrate the applicability of our analysis to the exponential case. The probability density function of the exponential distribution with parameter $\lambda$ is $f(t) = \lambda e^{-\lambda t}$, with $t \geq 0$. We deduce that $F^*(s) = \frac{\lambda}{\lambda+s}$, for $s \geq 0$. Hence, after some algebra, we get $q = \frac{\lambda(\lambda+\mu_2)}{\mu_2(\lambda+\mu_1)}$ under the stability condition $\lambda < \sqrt{\mu_1\mu_2}$. We thus deduce that

$$\begin{aligned}
\pi_{x,2} &= \pi_{0,2}\frac{\lambda^2}{\mu_2(\mu_1+\lambda)}\left(\frac{\lambda(\lambda+\mu_2)}{\mu_2(\lambda+\mu_1)}\right)^{x-1}, \text{ for } x \geq 1, \\
\pi_{x,1} &= \frac{\mu_2}{\lambda}\pi_{0,2}\frac{\lambda^2}{\mu_2(\mu_1+\lambda)}\left(\frac{\lambda(\lambda+\mu_2)}{\mu_2(\lambda+\mu_1)}\right)^{x-1}, \text{ for } x \geq 1, \text{ and,} \\
\pi_{0,2} &= \frac{\mu_1\mu_2 - \lambda^2}{\mu_2(\lambda+\mu_1)}.
\end{aligned}$$

We also have

$$\mu = \frac{\mu_2(\lambda+\mu_1)}{\lambda+\mu_2}, \text{ and } E(W) = \frac{\lambda}{\mu(\mu-\lambda)}.$$

We note that the expression of $E(W)$ only depends on $\mu$. Hence in the exponential case, having two different service rates does not provide improvement beyond what can be obtained with the average service rate $\mu$.

# 5 Analysis with Contract 1

In this section, we analyze Problem 1 with Contract 1. Contract 1 is simpler to analyze than Contract 2 since only one service rate has to be selected by the agent which makes the problem one-dimensional. In Proposition 1, we prove that the agent's utility is concave in $\mu_2$ when $\mu_1$ is kept constant. This shows that for each value of the piece-rate $pr$, there exists a unique $\mu_2$ which maximizes the agent's utility. Having a unique maximum of the utility is interesting as it shows that the agent will not select a service rate that is not globally optimal.

We prove this result by showing that $E(W)$ is decreasing and convex in $\mu_2$. This result could be expected from Weber (1983) who proved that the expected waiting time is decreasing and convex in the service rate for single server queues. However, the queueing model in Weber (1983) assumes that arrival times and service times are independent. This is not the case with event-dependent service rates. Further, Tu and Kumin (1983) showed that the convexity property does not hold without the independence of the two processes. Therefore, the convexity property of $E(W)$ has to be proven for our model. Note that even if the result might seem expected, we could possibly have found a counterexample. For instance, although expected from an M/M/1 queue, the expected waiting time in the G/M/1 is not convex in the mean arrival rate. A counterexample can be found in Fridgeirsdottir and Chiu (2005) with a Bernoulli interarrival time.

Further, we prove that $\mu_2 \geq \mu_1$. This result is expected as it is the purpose of the incentive $pr$ to speed up the agent's service. Finally, we also show that the optimal piece-rate $pr$, given by

$$pr = \frac{\left(e - (e-b)\frac{\partial \mu}{\partial \mu_2}\right) E(A)}{-\frac{\partial E(W)}{\partial \mu_2}}, \tag{13}$$

is positive and increasing in $\mu_2$ and $\mu$. The fixed-wage $FW$ compensates the base-effort but does not create an incentive to speed up service. Therefore, the principal should set $FW = b\mu$ in order to minimize the agent's fixed payout. The principal's cost, $C$, of inducing service rate $\mu$ is then given by

$$C = b\mu + \frac{\left(e - (e-b)\frac{\partial \mu}{\partial \mu_2}\right) (E(W_e) - E(W))}{-\frac{\partial E(W)}{\partial \mu_2}}.$$

Proposition 1 also proves that $C$ is increasing in $\mu_2$ and $\mu$. Having $pr$ and $C$ increasing in the obedient effort $\mu$ is also expected as the principal should increase the agent's payout if an higher effort is expected.

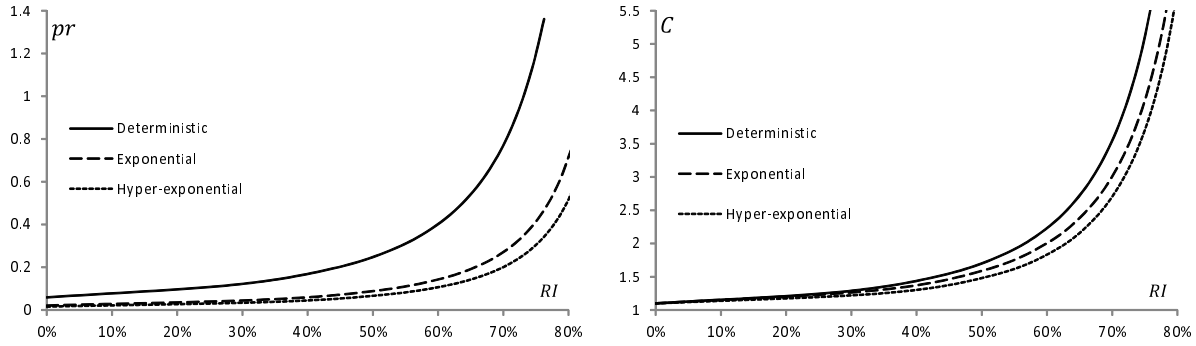
**Proposition 1.** *With Contract 1, the agent's problem is concave in $\mu_2$ and the service rate $\mu_2$ is selected by the agent such that $\mu_2 \geq \mu_1$. Further, the optimal piece-rate $pr$ and the agent's payout $C$ are positive and increasing in $\mu_2$ and $\mu$.*

In Figure 1, we present the optimal piece-rate $pr$ and the agent's payout $C$ as functions of the relative improvement, $RI$, obtained in $E(W)$, computed as $RI = \frac{E(W_e)-E(W)}{E(W_e)}$. We present the

cases of a deterministic, exponential and hyper-exponential distributions with the same expected inter-arrival time $E(A) = 1$. The hyper-exponential distribution is defined with two rates, 2/3 and 2, and a probability of 50% to observe an inter-arrival time with one of the two rates. Using (9), we show that $\lim_{\mu_2 \to \infty} q = F^*(\mu_1)$ and $\lim_{\mu_2 \to \infty} \mu = \frac{\mu_1}{1-F^*(\mu_1)}$. Therefore, we have $\lim_{\mu_2 \to \infty} E(W) = \frac{F^*(\mu_1)}{\mu_1}$. This asymptotic result gives an upper bound for $RI$ since $E(W)$ is decreasing in $\mu_2$. Starting from a situation where $\mu_2 = \mu_1$ and where $q = q_e$ is given by $q_e = F^*(\mu_1(1-q_e))$ as for a standard G/M/1 queue, the upper bound for $RI$ is equal to $1 - \frac{q_e}{\mu_1(1-q_e)}$. For the distributions presented in Figure 1, $RI$ can go up to 80%.

As expected from Proposition 1, $pr$ and $C$ are increasing in the relative improvement on $E(W)$. We also observe that these costs are convex in $RI$. As compared to the initial situation where $RI = 0\%$, the cost of increasing the obedient effort by incentives can grow extremely high if a high improvement is expected. However, when $RI$ is less than 30%, the agent's payout $C$ only increases of at most 20%. Therefore, incentives should be employed only when the principal expects a limited improvement as compared to the base case $\mu_1 = \mu_2 = \mu_e$. We also observe that the variability of the inter-arrival time distribution impacts the cost of increasing the obedient effort in the sense that this cost increases when the variability of the inter-arrival time decreases. Owing to the convexity of the expected waiting time in $\mu_2$, the effect of increasing $\mu_2$ is stronger when $E(W)$ is already high. Given that $E(W)$ increases with the variability of the inter-arrival time, increasing $\mu_2$ has more effect on $E(W)$ for the hyper-exponential distribution than for the deterministic one. That is, $-\frac{\partial E(W)}{\partial \mu_2}$ increases with the variability of the inter-arrival time. Since $-\frac{\partial E(W)}{\partial \mu_2}$ is at the denominator of the optimal piece-rate $pr$ (see (13)), the piece-rate and the agent's payout decrease with the inter-arrival time variability.

With Contract 1, the service rate $\mu_1$ is contractible. For the principal, with the aim of achieving a given service level on $E(W)$, the question is to determine how the value of $\mu_1$ should be set in the contract. For the principal, the selection of $\mu_1$ is a way to decide for the proportion given to the variable part of the agent's payout in comparison with the fixed part. Our numerical investigations indicate that it is advisable to have $\mu_1$ as close as possible to $\mu_2$, such that the average service rate becomes almost fully contractible and the incentive part of the payout is reduced to zero. This result was expected from contract theory. A system where the agent's actions are fully contractible is called a first-best setting. From Laffont and Martimort (2009), the first-best setting serves as a




(a) Optimal piece-rate $pr$ (b) Agent's payout $C$

Figure 1: Numerical illustration ($e = b = 1$, $\mu_1 = 1.1$, $E(A) = 1$)

lower bound for the agent's payout.

Although less costly, a fully contractible system is less satisfying on a human resource management level. Due to the strong convexity of $pr$ and $C$ in $RI$, our analysis reveals that the agent's payout can be kept relatively close to the first-best cost for a large range of relative improvement. The analysis of the first contract showed that having different service rates could be a way to increase the average service rate while having the agent's actions partially contractible. In the following section, we further investigate the benefits of having different service rates on the expected waiting time by assuming that the average service rate $\mu$ is contractible and that the agent has full discretion for selecting $\mu_1$ and $\mu_2$.

## 6 Analysis with Contract 2

With Contract 2, the average service rate $\mu$ is contractible and kept constant but the two service rates $\mu_1$ and $\mu_2$ are optimized. Therefore, for a given value of $\mu$, $\mu_1$ and $\mu_2$ are related via

$$\mu_2 = \frac{\mu_1 F^*(\mu_1)\mu}{\mu_1 - \mu(1 - F^*(\mu_1))}. \tag{14}$$

Using this relation, we first prove in Lemma 2 that $\mu_2$ is decreasing in $\mu_1$.

**Lemma 2.** *The service rate $\mu_2$ is decreasing in $\mu_1$ when $\mu$ is held constant.*

Next, we question whether having two different service rates has the potential to reduce the expected waiting time. In particular, we question whether the agent should accelerate after an arrival ($\mu_1 > \mu_2$) or after a service completion ($\mu_2 > \mu_1$) while keeping the average service rate

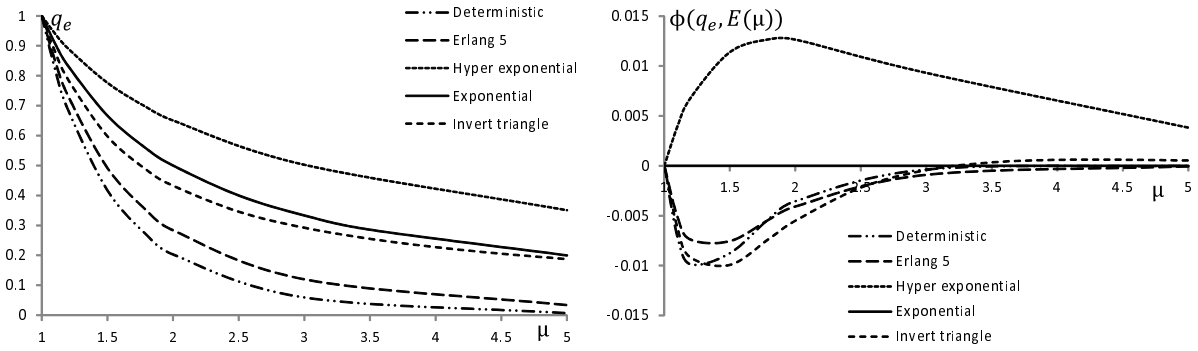
fixed.

In Proposition 2, we specify how the expected waiting time behaves when either $\mu_1$ or $\mu_2$ tends to infinity for any distribution of the inter-arrival time. The result of Proposition 2 also shows that with Contract 2, $\mu_2 = \infty$ cannot be optimal since the $G/M^{event}/1$ queue behaves as a standard $G/M/1$ queue with equal service rates. Moreover, if the variability of the inter-arrival time is lower than the one of an exponential distribution, $\mu_1 = \infty$ can also not be optimal as the $G/M^{event}/1$ behaves as an $M/M/1$ queue in this asymptotic case. It should be noted that even when the variability of the inter-arrival time is higher than the one of an exponential distribution, $\mu_1 = \infty$ is not optimal (see Theorem 2). This result shows that the optimization of event-dependent service rates differs from the optimization of state-dependent service rates. Recall that when the service rate can be dynamically adjusted to the system size, the optimal rates are located on the boundary of their value domains (that is, a bang-bang control) for optimization problems where a trade-off between a holding cost and a mean service rate cost has to be determined (Ma and Ao, 1994; Kumar et al., 2013; Xia, 2014; Xia et al., 2017).

**Proposition 2.** *When the expected service rate, $\mu$, is held constant, then:*

- *As $\mu_1$ tends to infinity, $q$ tends to $\frac{1}{\mu E(A)}$. This means that the queue behaves as in the case where the inter-arrival time is exponentially distributed.*
- *As $\mu_2$ tends to infinity, $q$ tends to $q_e$. This means that the queue behaves as in the case with equal service rates (that is, $\mu_1 = \mu_2 = \mu_e$).*

Using Proposition 2, we prove in Theorem 2 that $E(W)$ has a unique minimum in $\mu_1$ or $\mu_2$ and we provide a criterion for having this minimum with $\mu_1 > \mu_2$ or $\mu_2 < \mu_1$. This criterion also indicates whether the agent should accelerate after a service completion or arrival instant to maximize their utility. Showing that $\mu_1 = \mu_2$ is not optimal proves that the expected waiting time is not only driven by the mean service rate. Further, this shows that the optimal solution in this study differs from the one in other queueing models where a server alternates between different service rates. For instance in the case where a server alternates between $\mu_1$ and $\mu_2$ from one customer to another in a cyclic way, if the mean service time is kept constant then the expected wait is minimized for $\mu_1 = \mu_2$ (Zhou and Gans, 1999).




(a) $q_e$ as a function of $\mu$

(b) $\phi(q_e, \mu)$ as a function of $\mu$

Figure 2: Effect of $\mu$ on $q_e$ and $\phi(q_e, \mu)$ $(E(A) = 1)$

**Theorem 2.** *We define the expression* $\phi(q_e, \mu)$ *as*

$$\phi(q_e, \mu) = q_e - F^*(\mu) + q_e\mu(1 - F^*(\mu))\frac{\partial F^*}{\partial s}|_{s=\mu(1-q_e)},$$

*where* $q_e$ *is the unique solution of* $q_e = F^*(\mu(1 - q_e))$ *(i.e., the value of* $q$ *when* $\mu_1 = \mu_2 = \mu$*). The expected waiting time* $E(W)$ *has a unique minimum in* $\mu_1$ *or in* $\mu_2$ *if* $\phi(q_e, \mu) \neq 0$*. Further,*

- *If* $\phi(q_e, \mu) > 0$*, then* $E(W)$ *is minimized for a couple* $(\mu_1, \mu_2)$ *with* $\mu_1 > \mu_2$*,*
- *If* $\phi(q_e, \mu) < 0$*, then* $E(W)$ *is minimized for a couple* $(\mu_1, \mu_2)$ *with* $\mu_1 < \mu_2$*.*

In Figure 2, we present $q_e$ and $\phi(q_e, \mu)$ as functions of $\mu$ for different inter-arrival time distributions. We adjust the parameters of each distribution such that the expected inter-arrival time is equal to 1. We consider the deterministic distribution with $f(t) = \delta_{t=1}$ (where $\delta$ is the Dirac function), the Erlang 5 distribution with $f(t) = \frac{5^5 t^4 e^{-5t}}{4!}$, the hyper-exponential distribution with $f(t) = \frac{\frac{4}{7}e^{-\frac{4}{7}t} + e^{-4t}}{2}$, the exponential distribution with $f(t) = e^{-t}$, and the invert triangle distribution with $f(t) = 1 - t$ for $0 \leq t \leq 1$, $f(t) = t - 1$ for $1 \leq t \leq 2$, and $f(t) = 0$, for $t > 2$. Figure 2(a) is presented to show how $q_e$ evolves as a function of $\mu$ for the considered distributions. As expected, $q_e$ and the expected waiting time are decreasing with $\mu$ and increasing with the variability of the inter-arrival time distribution. Figure 2(b) presents $\phi(q_e, \mu)$ as a function of $\mu$. For the exponential distribution, we find that $\phi(q_e, \mu) = 0$ which indicates that we cannot find a couple $(\mu_1, \mu_2)$ with a lower waiting time than in the case $\mu_1 = \mu_2 = \mu$. For a distribution with a higher variability like the hyper-exponential distribution, we observe that $\phi(q_e, \mu) > 0$ which indicates that an improvement can be obtained when accelerating the speed of service after an arrival. On the contrary, for other

distributions with a lower variability than the exponential distribution like the Erlang 5 or the deterministic distribution, we have $\phi(q_e, \mu) < 0$ which suggests that the server should accelerate after a service completion. However, the expression of $\phi(q_e, \mu)$ in Theorem 2 cannot be reduced to the effect of the variability of the inter-arrival time. A counterexample is given with the invert triangle distribution for which $\phi(q_e, \mu)$ can be positive or negative depending on the value of $\mu$. The expression of $\phi(q_e, \mu)$ also does not reduce to having an increasing or decreasing failure rate property. For instance, the hyper-exponential distribution with density function $f(t) = p\lambda e^{-\lambda t} + (1-p)\delta_{t=0}$ has the increasing failure rate property but we do not have $\phi(q_e, \mu) > 0$.

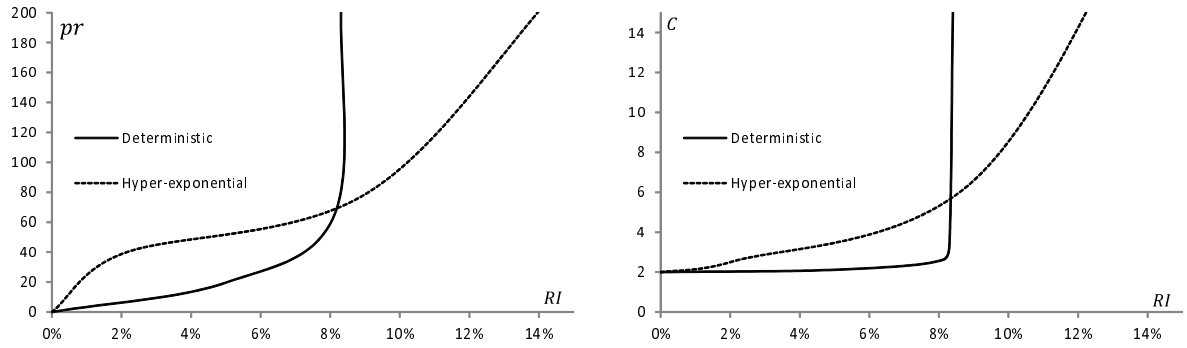
Consequently, the condition in Theorem 2 does not reduce to a simpler condition to estimate whether the agent should accelerate after a service completion or arrival instant. However in many cases, the variability of the inter-arrival time may serve as an indicator whether $E(W)$ can be minimized with $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$. Specifically, for most distributions with a higher (lower) variability than the exponential distribution, the agent should select $\mu_1$ and $\mu_2$ such that $\mu_1 > \mu_2$ ($\mu_1 < \mu_2$).

We now focus of the compensation parameters. The fixed-wage $FW$ compensates the base-effort but does not provide an incentive to modify the service rates. Therefore, the principal should set $FW = b\mu$. Next, Theorem 2 determines whether the agent selects $\mu_1 > \mu_2$ or $\mu_2 > \mu_1$. The optimal piece-rate compensation, $pr$, is then determined with

$$
\begin{aligned}
&pr = \frac{-eE(A)\frac{\partial \mu_2}{\partial \mu_1}}{\frac{\partial E(W)}{\partial \mu_1}}, \text{ if } \mu_1 < \mu_2, \text{ and, } pr = \frac{-eE(A)\frac{\partial \mu_1}{\partial \mu_2}}{\frac{\partial E(W)}{\partial \mu_2}}, \text{ if } \mu_2 < \mu_1, \text{ and} \qquad (15)\\
&C = b\mu - \frac{e\frac{\partial \mu_2}{\partial \mu_1}\left(E(W_e) - E(W)\right)}{\frac{\partial E(W)}{\partial \mu_1}}, \text{ if } \mu_1 < \mu_2, \text{ and,}\\
&C = b\mu - \frac{e\frac{\partial \mu_1}{\partial \mu_2}\left(E(W_e) - E(W)\right)}{\frac{\partial E(W)}{\partial \mu_2}}, \text{ if } \mu_2 < \mu_1.
\end{aligned}
$$

Starting from $\mu_1 = \mu_2 = \mu$, the minimum of $E(W)$ can only be reached if $pr = \infty$ since we have $\frac{\partial E(W)}{\partial \mu_1} = 0$ at the minimum of $E(W)$. Therefore, the value of $\mu_1$ ($\mu_2$) which minimizes $E(W)$ serves as a lower bound for the optimal value of $\mu_1$ ($\mu_2$) in the case $\mu_1 < \mu_2$ ($\mu_1 > \mu_2$). We deduce that $pr > 0$ since $\frac{\partial \mu_2}{\partial \mu_1} \leq 0$ (Lemma 2) and $\frac{\partial E(W)}{\partial \mu_1} > 0$ if $\mu_1 < \mu_2$ or $\frac{\partial E(W)}{\partial \mu_2} > 0$ if $\mu_1 > \mu_2$ (Theorem 2).

In Figure 3, we give the piece-rate $pr$ and the agent's payout $C$ as functions of the relative improvement obtained for the expected waiting time $E(W)$. We present the same deterministic




(a) Optimal piece-rate $pr$

(b) Agent's payout $C$

Figure 3: Numerical illustration ($e = b = 1$, $\mu = 2$, $E(A) = 1$)

and hyper-exponential inter-arrival time distributions as those considered in Figure 1. These distributions are representative of high and low variability distributions. The first observation is that when $\mu$ is held constant, the relative improvement obtained by selecting different service rates is more limited than with Contract 1. For the deterministic distribution $RI$ can go up to 8.2% and for the hyper-exponential distribution up to 15%. This shows that the average service rate is the main driver to reduce the expected waiting time. Therefore, having a fixed average service rate reduces the range of achievable improvement. It should also be noted that the range of achievable improvements increases with the variability of the inter-arrival time. This makes a difference with Contract 1 where the upper bound of $RI$ was identical for all distributions. This observation is related to the fact that $E(W)$ is no longer monotonous in $\mu_1$ and $\mu_2$.

As with Contract 1, improving the service level is costly to the principal, therefore the piece-rate $pr$ and agent's payout $C$ are increasing in the obedient effort, $RI$. However, the behavior of $pr$ and $C$ shows some differences as compared to the one with Contract 1. In particular, the convexity property does not hold with Contract 2. With an hyper-exponential distribution, the first part of the curves is concave, which indicates that a little improvement in the expected waiting time may lead to a high increase in cost. On the contrary, for the deterministic distribution, the agent's payout is almost insensitive to $RI$ when $RI$ is less than 6%. This suggests that it is advisable to select different service rates after an arrival or a service completion only for inter-arrival time distributions that have a low variability and when the expected improvement is limited.

# 7 Other benefits of selecting event-dependent service rates

In this section, we restrict the study to the exponential inter-arrival time distribution to determine whether having $\mu_1 \neq \mu_2$ could be beneficial for other performance measures than the expected waiting time. In Section 7.1, we evaluate the expected waiting time at arrival and in Section 7.2 we include the feature of abandonment.

## 7.1 Expected waiting time at arrival

When $\mu_1 = \mu_2 = \mu$, the expected waiting time at arrival when $x$ customers are already present in the system is $\frac{x}{\mu}$. This expression changes when $\mu_1 \neq \mu_2$. We denote by $H_x^1$, and $H_x^2$, the Laplace transform in the variable $s$, of the distribution function of the first passage time from state $(x, 1)$ and state $(x, 2)$ to a beginning of service, where states $(x, 1)$ and $(x, 2)$ correspond to situations where $x$ customers are present in the system after an arrival, and after a service completion, respectively. Applying the first-step analysis to the related discrete-time Markov chain (for instance, see Kulkarni (2016), p.162), we obtain the following finite set of equations:

$$\begin{aligned} &H_x^1(\mu_1 + s) = \mu_1 H_{x-1}^2, \text{ for } x \geq 1, \\ &H_x^2(\lambda + \mu_2 + s) = \mu_2 H_{x-1}^2 + \lambda H_x^1, \text{ for } x \geq 1, \text{ with} \\ &H_0^2 = 1. \end{aligned} \tag{16}$$

This leads to

$$H_x^1 = \frac{\mu_1}{\mu_1 + s} \left( \frac{\mu_2(\mu_1 + s) + \lambda\mu_1}{(\mu_1 + s)(\lambda + \mu_2 + s)} \right)^{x-1}, \text{ for } x \geq 1. \tag{17}$$

The expected waiting time at arrival, when $x$ customers are already present in the system, $E(W_x)$, is given by $E(W_x) = -\frac{\partial H_x^1}{\partial s}|_{s=0}$. Using (14) and (17), we deduce that

$$E(W_x) = \frac{(\lambda + \mu_1)x + \mu_2 - \mu_1}{\mu_1(\mu_2 + \lambda)} = \frac{(x-1)(\mu_1 - \mu) + \lambda x}{\lambda\mu_1}, \text{ for } x \geq 1.$$

This expression is increasing in $\mu_1$ if and only if $x \geq \frac{\mu}{\mu - \lambda}$. This means that having $\mu_1 > \mu_2$ reduces short waiting times while increasing long waiting times. The opposite is true when $\mu_1 < \mu_2$. This conclusion can be explained intuitively. Recall that in the case of an exponential inter-arrival

time, the expected waiting time is insensitive to the service rate differentiation. Therefore, if some customers benefit from $\mu_1 < \mu_2$, it should be detrimental to others. When $\mu_1 < \mu_2$, an arriving customer will have to wait a long time for the customer in service to be served as the arriving customer has just generated an arrival event. Once the customer in service is served (if no other arrival occurred), the following services will be fast. The effect of having fast services is beneficial for the arriving customer if there are many customers to be served. That is when the queue size is long. When the queue size is short, the detrimental effect of waiting a long time for the customer in service to be served is not compensated by following fast services.

In the asymptotic case where $\mu_1$ tends to $\mu - \lambda$ (that is, when $\mu_2$ tends to infinity), the expected waiting time at arrival becomes insensitive to the number of customers present in the queue. In this case, either an arriving customer finds the agent available and there is no wait, or there is at least one customer present in the system and the expected waiting time is equal to $\frac{1}{\mu-\lambda}$, which corresponds to the expected time spent in an M/M/1 queue. In practice, it may not be possible to achieve an infinite value for $\mu_2$. However, increasing the value of $\mu_2$ results in reducing the sensitivity of the expected waiting time at arrival to the number of customers present. This can be interesting if the principal cares about fairness among customers.

### 7.2 Abandonment

We now include the feature of abandonment in the model description. We assume that customers in the queue have a limited patience that is exponentially distributed with rate $\beta$ and which does not impact the agent's service rate selection. The system balance equations are given by

$$
\begin{aligned}
&\lambda\pi_{0,2} = \mu_1\pi_{1,1} + \mu_2\pi_{1,2}, \qquad (18)\\
&(\lambda + \mu_2 + \beta(x-1))\pi_{x,2} = (\mu_1 + \beta x)\pi_{x+1,1} + (\mu_2 + \beta x)\pi_{x+1,2}, \text{ for } x \geq 1,\\
&(\lambda + \mu_1)\pi_{1,1} = \lambda\pi_{0,2}, \text{ and,}\\
&(\lambda + \mu_1 + \beta(x-1))\pi_{x,1} = \lambda\pi_{x-1,2} + \lambda\pi_{x-1,1}, \text{ for } x \geq 2.
\end{aligned}
$$

After solving (18), we obtain

$$\pi_{x,1} = \pi_{0,2}\frac{\left(\frac{\lambda}{\beta}\right)^x \Gamma\left(\frac{\lambda+\mu_2}{\beta}+x-1\right)}{\Gamma\left(\frac{\lambda+\mu_1}{\beta}+x\right)\Gamma\left(\frac{\mu_2}{\beta}+x-1\right)}\frac{\Gamma\left(\frac{\lambda+\mu_1}{\beta}\right)\Gamma\left(\frac{\mu_2}{\beta}\right)}{\Gamma\left(\frac{\lambda+\mu_2}{\beta}\right)}, \text{ for } x \geq 1,$$

$$\pi_{x,2} = \pi_{0,2}\frac{\left(\frac{\lambda}{\beta}\right)^{x+1} \Gamma\left(\frac{\lambda+\mu_2}{\beta}+x-1\right)}{\Gamma\left(\frac{\lambda+\mu_1}{\beta}+x\right)\Gamma\left(\frac{\mu_2}{\beta}+x\right)}\frac{\Gamma\left(\frac{\lambda+\mu_1}{\beta}\right)\Gamma\left(\frac{\mu_2}{\beta}\right)}{\Gamma\left(\frac{\lambda+\mu_2}{\beta}\right)}, \text{ for } x \geq 1, \text{ and,}$$

$$\pi_{0,2} = \left[\sum_{x=0}^{\infty}\frac{\left(\frac{\lambda}{\beta}\right)^x \Gamma\left(\frac{\lambda+\mu_2}{\beta}+x\right)}{\Gamma\left(\frac{\lambda+\mu_1}{\beta}+x\right)\Gamma\left(\frac{\mu_2}{\beta}+x\right)}\frac{\Gamma\left(\frac{\lambda+\mu_1}{\beta}\right)\Gamma\left(\frac{\mu_2}{\beta}\right)}{\Gamma\left(\frac{\lambda+\mu_2}{\beta}\right)}\right]^{-1},$$

where the gamma function, $\Gamma(z)$, is defined as $\Gamma(z) = \int_{t=0}^{\infty} t^{z-1}e^{-t}\mathrm{d}t$, with $z > 0$. We deduce that the probability of abandonment, $P_A$, is given by

$$P_A = \frac{\beta}{\lambda}\frac{\sum_{x=1}^{\infty}(x-1)\frac{\left(\frac{\lambda}{\beta}\right)^x\Gamma\left(\frac{\lambda+\mu_2}{\beta}+x\right)}{\Gamma\left(\frac{\lambda+\mu_1}{\beta}+x\right)\Gamma\left(\frac{\mu_2}{\beta}+x\right)}}{\sum_{x=0}^{\infty}\frac{\left(\frac{\lambda}{\beta}\right)^x\Gamma\left(\frac{\lambda+\mu_2}{\beta}+x\right)}{\Gamma\left(\frac{\lambda+\mu_1}{\beta}+x\right)\Gamma\left(\frac{\mu_2}{\beta}+x\right)}}.$$

Using these expressions, we investigate the effect of having $\mu_1 \neq \mu_2$ on the fraction of lost customers. In Table 2, we present different combinations of the parameters $\lambda$, $\mu_1$, and $\mu_2$, to reflect various situations of workload and gap between $\mu_1$ and $\mu_2$. For each situation, we compute the probability to have an empty system, $\pi_{0,2}$, and the equivalent service rate $\mu_e$ which would lead to the same value of $\pi_{0,2}$, with $\mu_e = \mu_1 = \mu_2$. In this way, we obtain two systems, one with unequal and one with equal service rates, which can be compared in terms of fraction of lost customers, $P_A$ and $P_A^e$, respectively. In the last column, we compute the difference in terms of lost customers, computed as D= $P_A^e - P_A$.

Table 2: Effect of having different service rates on the proportion of abandonment ($\beta = 2$)

| $\mu_1$ | $\mu_2$ | $\lambda$ | $\pi_{0,2}$ | $\mu_e$ | $P_A$ | $P_A^e$ | D |
|---|---|---|---|---|---|---|---|
| 2 | 0.5 | 1 | 42.81% | 1.050 | 34.20% | 39.93% | 5.73% |
| 2 | 0.5 | 2 | 17.49% | 0.883 | 55.17% | 63.57% | 8.40% |
| 2 | 0.5 | 4 | 3.87% | 0.781 | 73.28% | 81.24% | 7.96% |
| 0.5 | 2 | 1 | 40.89% | 0.978 | 44.87% | 42.19% | -2.68% |
| 0.5 | 2 | 2 | 23.43% | 1.194 | 58.08% | 54.28% | -3.80% |
| 0.5 | 2 | 4 | 8.34% | 1.392 | 72.26% | 68.10% | -4.16% |

We observe that having $\mu_1 > \mu_2$ reduces the fraction of lost customers as compared to a system with equal rates while the opposite effect occurs when $\mu_1 < \mu_2$. As expected, we also observe that

the effect of having different service rates is stronger in congested systems (that is, with a high arrival rate). Having $\mu_1 > \mu_2$ results in an agent being fast in periods of queue size growth while being slow when the queue size reduces. The opposite is true when $\mu_1 < \mu_2$. This means that the transitions between an empty system and a system with a high congestion are made faster and more frequently when $\mu_1 < \mu_2$ than when $\mu_2 < \mu_1$. Thus, a system with $\mu_1 < \mu_2$ will regularly be highly congested, resulting in a high fraction of lost customers as shown in Table 2. Therefore, accelerating the speed of service after an arrival can be an efficient strategy to reduce the fraction of lost customers.

# 8 Conclusion

We investigated a principal-agent problem for a G/M/1 queue where the principal rewards the agent with a performance-based contract. The agent adjusts the service rate to the last realized event being an arrival or a service completion with a utility-maximizer objective. We studied this problem for two contracts. In the first one, the service rate after an arrival is contractible whereas in the second one the average service rate is contractible. With the first contract, we proved that the agent's utility is concave in the obedient effort. This allowed us to express the compensation parameters which minimize the agent's payout. Through incentives, the range of achievable improvement in the expected waiting time is very large with the first contract. However, the convexity of the agent's payout in the realized improvement indicates that the cost of a contract with incentives can grow too high if the desired improvement is high. With the second contract, we proved that the expected waiting time has a unique minimum and we presented a criterion indicating whether accelerating after an arrival or after a service completion was optimal. Next, we expressed the payout parameters and observed that the range of achievable improvement is less with the second contract than with the first one. Finally, with an exponential inter-arrival time, we determined other positive aspects of having event-dependent service rates. Accelerating after a service completion is proven to provide more fairness among arriving customers while accelerating after an arrival reduces the fraction of lost customers due to abandonment.

This opens up several avenues for future research. It would be interesting to include other features corresponding to customer behavior such as retrial, reconnect or workload-dependency in their arrival process. Another extension of the model concerns the possibility for the agent to

take decisions not just based on the last event but on a larger set of past events. We could also consider a non-exponential distribution for the service time. Losing the memoryless property of the exponential distribution would make the model description more difficult, however, especially with regard to defining the remaining service time when a new event occurs. We could also extend the analysis to a multi-server setting. The analysis of the G/M/s queue can be made in a similar way as the one of the G/M/1 queue. However, it may become complex to make assumptions on the agents' behavior when a service completion occurs. We may need to consider from which agent the service completion happened. Moreover, in the utility-maximizer perspective of the agents, we may also need to make assumptions whether agents collaborate or not for service rate optimization.

## References


Adan, I., Boxma, O., and Perry, D. (2005). The G/M/1 queue revisited. *Mathematical Methods of Operations Research*, 62(3):437–452.

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

Avi-Itzhak, B., Golany, B., and Rothblum, U. G. (2006). Strategic equilibrium versus global optimum for a pair of competing servers. *Journal of Applied Probability*, 43(4):1165–1172.

Bae, J. and Kim, S. (2010). The stationary workload of the G/M/1 queue with impatient customers. *Queueing Systems*, 64(3):253–265.

Baiman, S., Netessine, S., and Saouma, R. (2010). Informativeness, incentive compensation, and the choice of inventory buffer. *The Accounting Review*, 85(6):1839–1860.

Cachon, G. and Zhang, F. (2007). Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science*, 53(3):408–420.

Chae, K., Lee, S., and Lee, H. (2006). On stochastic decomposition in the GI/M/1 queue with single exponential vacation. *Operations Research Letters*, 34(6):706–712.

Chan, C., Yom-Tov, G., and Escobar, G. (2014). When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482.

Chen, F., Lai, G., and Xiao, W. (2016). Provision of incentives for information acquisition: Forecast-based contracts vs. menus of linear contracts. *Management Science*, 62(7):1899–1914.

Christ, D. and Avi-Itzhak, B. (2002). Strategic equilibrium for a pair of competing servers with convex cost and balking. *Management Science*, 48(6):813–820.

Dai, T., Ke, R., and Ryan, C. (2021). Incentive design for operations-marketing multitasking. *Management Science*, 67(4):2211–2230.

Delasay, M., Ingolfsson, A., Kolfal, B., and Schultz, K. (2019). Load effect on service times. *European Journal of Operational Research*, 279(3):673–686.

Fellner, G. and Sutter, M. (2009). Causes, consequences, and cures of myopic loss aversion - An experimental investigation. *The Economic Journal*, 119(537):900–916.

Fridgeirsdottir, K. and Chiu, S. (2005). A note on convexity of the expected delay cost in single-server queues. *Operations Research*, 53(3):568–570.

Geng, X., Huh, W., and Nagarajan, M. (2015). Fairness among servers when capacity decisions are endogenous. *Production and Operations Management*, 24(6):961–974.

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

Gopalakrishnan, R., Doroudi, S., Ward, A., and Wierman, A. (2016). Routing and staffing when servers are strategic. *Operations Research*, 64(4):1033–1050.

Guo, P., Tang, C., Wang, Y., and Zhao, M. (2019). The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: Fee-for-service versus bundled payment. *Manufacturing & Service Operations Management*, 21(1):154–170.

Hassin, R. and Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media.

Haviv, M. and Kerner, Y. (2011). The age of the arrival process in the G/M/1 and M/G/1 queues. *Mathematical Methods of Operations Research*, 73(1):139–152.

Herweg, F., Müller, D., and Weinschenk, P. (2010). Binary payment schemes: Moral hazard and loss aversion. *American Economic Review*, 100(5):2451–77.

Hokstad, P. (1975). The G/M/m queue with finite waiting room. *Journal of Applied Probability*, pages 779–792.

Jain, S. (2012). Self-control and incentives: An analysis of multiperiod quota plans. *Marketing Science*, 31(5):855–869.

Jiang, H., Pang, Z., and Savin, S. (2012). Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management*, 14(4):654–669.

Kalai, E., Kamien, M., and Rubinovitch, M. (1992). Optimal service speeds in a competitive environment. *Management Science*, 38(8):1154–1163.

Ke, J. and Wang, K. (2002). A recursive method for the N policy G/M/1 queueing system with finite capacity. *European Journal of Operational Research*, 142(3):577–594.

Kerner, Y. (2008). The conditional distribution of the residual service time in the $M_n$/G/1 queue. *Stochastic Models*, 24(3):364–375.

Kleinrock, L. (1975). *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication.

Kulkarni, V. (2016). *Modeling and analysis of stochastic systems*. Crc Press.

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

Laffont, J. and Martimort, D. (2009). *The theory of incentives: The principal-agent model*. Princeton University press.

Lal, R. and Srinivasan, V. (1993). Compensation plans for single-and multi-product salesforces: An application of the Holmstrom-Milgrom model. *Management Science*, 39(7):777–793.

Laslett, G. (1975). Characterising the finite capacity GI/M/1 queue with renewal output. *Management Science*, 22(1):106–110.

Legros, B. (2021). Age-based Markovian approximation of the G/M/1 queue. *Operations Research Letters*, 49(5):708–714.

Legros, B. and Sezer, D. (2018). Stationary analysis of a single queue with remaining service time-dependent arrivals. *Queueing Systems*, 88(1-2):139–165.

Li, S., Chen, K., and Rong, Y. (2020). The behavioral promise and pitfalls in compensating store managers. *Management Science*, 66(10):4899–4919.

Long, X. and Nasiry, J. (2020). Wage transparency and social comparison in sales force compensation. *Management Science*, 66(11):5290–5315.

Löpker, A. and Perry, D. (2010). The idle period of the finite G/M/1 queue with an interpretation in risk theory. *Queueing Systems*, 64(4):395–407.

Ma, D. and Ao, X. (1994). A direct approach to decentralized control of service rates in a closed Jackson network. *IEEE Transactions on Automatic Control*, 39(7):1460–1463.

Oz, B., Adan, I., and Haviv, M. (2017). A rate balance principle and its application to queueing models. *Queueing Systems*, 87(1):95–111.

Pourbabai, B. (1990). Tandem behavior of a finite capacity G/M/1 queueing system: An algorithm. *European Journal of Operational Research*, 46(3):380–387.

Ren, Z. and Zhou, Y. (2008). Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383.

Suen, S., Negoescu, D., and Goh, J. (2018). Design of incentive programs for optimal medication adherence. *Available at SSRN 3308510*.

Tu, H. and Kumin, H. (1983). A convexity result for a class of GI/G/1 queueing systems. *Operations Research*, 31(5):948–950.

Weber, R. (1983). A note on waiting times in single server queues. *Operations Research*, 31(5):950–951.

Xia, L. (2014). Service rate control of closed Jackson networks from game theoretic perspective. *European Journal of Operational Research*, 237(2):546–554.

Xia, L., He, Q., and Alfa, A. (2017). Optimal control of state-dependent service rates in a MAP/M/1 queue. *IEEE Transactions on Automatic Control*, 62(10):4965–4979.

Zafer, M. and Modiano, E. (2005). A calculus approach to minimum energy transmission policies with quality of service guarantees. In *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 1, pages 548–559. IEEE.

Zhan, D. and Ward, A. (2018). The M/M/1+ M queue with a utility-maximizing server. *Operations Research Letters*, 46(5):518–522.

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

Zhang, Z. and Tian, N. (2004). The N threshold policy for the GI/M/1 queue. *Operations Research Letters*, 32(1):77–84.

Zhou, Y. and Gans, N. (1999). A single-server queue with Markov modulated service times. *submitted for publication.*

# Appendix

## A Proof of Lemma 1

*Proof.* Let us start with (3). Using the transition rates in the Markov chain during a duration $dt$, we obtain for $x \geq 1$,

$$p_{t+dt}(x,2,r) = (1-\mu_2 dt)p_t(x,2,r+dt) + \mu_2 dt p_t(x+1,2,r+dt) + \mu_1 dt p_t(x+1,1,r+dt).$$

Taking $t \longrightarrow \infty$ leads to

$$\frac{p(x,2,r+dt) - p(x,2,r)}{dt} = \mu_2 p(x,2,r+dt) - \mu_2 p(x+1,2,r+dt) + \mu_1 p(x+1,1,r+dt),$$

for $x \geq 1$. Next, taking $dt \longrightarrow 0$, we obtain (3). Equation (2) can be deduced from (3) by removing the term $\mu_2 p(x,2,r)$ as the agent is not active when the system is empty.

Consider now (5). The transition rates during a duration $dt$ leads to

$$p_{t+dt}(x,1,r) = (1-\mu_1 dt)p_t(x,1,r+dt) + f(r)dt p_t(x-1,1,0) + f(r)dt p_t(x-1,2,0),$$

for $x \geq 2$. Dividing this expression by $dt$, letting $t$ tend to infinity and next $dt$ tend to zero, leads to (5). Equation (4) can be deduced from (5) by removing the term $f(r)p(x-1,1,0)$ since the system cannot be empty after an arrival. □

## B Proof of Theorem 1

*Proof.* We now integrate both sides of (2)-(5) for $r$ from 0 to $\infty$. For the left hand side of the equations, we have $\int_{r=0}^{\infty} p(x,i,r)'\,\mathrm{d}r = \lim_{r\longrightarrow\infty} p(x,i,r) - p(x,i,0) = -p(x,i,0)$, for stability reasons. For the right hand side of the equations, we denote by $\pi_{x,i}$ the stationary probability to have $x$ customers in the system with the last event being an arrival (that is, $i=1$) or a service completion

(that is, $i = 2$). Given that $\int_{r=0}^{\infty} f(r)\,\mathrm{d}r = 1$, we obtain

$$-p(0,2,0) = -\mu_2\pi_{1,2} - \mu_1\pi_{1,1}, \tag{19}$$

$$-p(x,2,0) = \mu_2\pi_{x,2} - \mu_2\pi_{x+1,2} - \mu_1\pi_{x+1,1}, \text{ for } x \geq 1, \tag{20}$$

$$-p(1,1,0) = \mu_1\pi_{1,1} - p(0,2,0), \text{ and} \tag{21}$$

$$-p(x,1,0) = \mu_1\pi_{x,1} - (p(x-1,1,0) + p(x-1,2,0)), \text{ for } x \geq 2. \tag{22}$$

By summing up (20) and (22), we obtain

$$\mu_2\pi_{x+1,2} + \mu_1\pi_{x+1,1} - (p(x,2,0) + p(x,1,0)) = \mu_2\pi_{x,2} + \mu_1\pi_{x,1} - (p(x-1,2,0) + p(x-1,1,0)),$$

for $x \geq 2$. This means that the sequence $\mu_2\pi_{x+1,2} + \mu_1\pi_{x+1,1} - (p(x,2,0) + p(x,1,0))$ is constant for $x \geq 1$. For stability reason, the limit of the sequence is equal to zero as $x$ tends to infinity. This proves that

$$\mu_2\pi_{x+1,2} + \mu_1\pi_{x+1,1} = p(x,2,0) + p(x,1,0), \text{ for } x \geq 1. \tag{23}$$

Note that this equality is not valid for $x = 0$ as shown by (19).

From (23), we have $p(x-1,2,0) + p(x-1,1,0) = \mu_2\pi_{x,2} + \mu_1\pi_{x,1}$, for $x \geq 2$. Therefore (22) can be rewritten as

$$\mu_2\pi_{x,2} = p(x,1,0), \text{for } x \geq 2. \tag{24}$$

Note that this equality is also valid for $x = 1$. This can be shown by summing up (19) and (21).

We now determine a recursive way to compute the stationary probabilities. We define the LST of the probabilities $p(x,i,r)$ as $P^*(x,i,s) = \displaystyle\int_{r=0}^{\infty} e^{-sr}p(x,i,r)\,\mathrm{d}r$. In what follows, we relate the

probabilities $p(x,i,0)$ with the LSTs of $p(x,i,r)$ at $s=\mu_1$ and $s=\mu_2$:

$$p(0,2,0) = \mu_2 P^*(1,2,\mu_2) + \mu_2 P^*(0,2,\mu_2) + \mu_1 P^*(1,1,\mu_2), \tag{25}$$

$$p(x,2,0) = \mu_2 P^*(x+1,2,\mu_2) + \mu_1 P^*(x+1,1,\mu_2), \text{ for } x \geq 1, \tag{26}$$

$$p(1,1,0) = p(0,2,0)F^*(\mu_1), \text{ and} \tag{27}$$

$$p(x,1,0) = (p(x-1,2,0) + p(x-1,1,0))F^*(\mu_1), \text{ for } x \geq 2. \tag{28}$$

We explain here how (25)-(28) are computed. We start with (26). After multiplication by $e^{-\mu_2 r}$, (3) can be rewritten as

$$p(x,2,r)'e^{-\mu_2 r} - \mu_2 p(x,2,r)e^{-\mu_2 r} = -\mu_2 p(x+1,2,r)e^{-\mu_2 r} - \mu_1 p(x+1,1,r)e^{-\mu_2 r}, \text{ for } x \geq 1.$$

Hence,

$$(p(x,2,r)e^{-\mu_2 r})' = -\mu_2 p(x+1,2,r)e^{-\mu_2 r} - \mu_1 p(x+1,1,r)e^{-\mu_2 r}, \text{ for } x \geq 1.$$

Integrating this equation for $r$ from 0 to infinity yields (26). Equation (25) can be deduced from (26) by adding the corrective term $\mu_2 P^*(0,2,\mu_2)$. With the same approach, (5) can be rewritten as

$$(p(x,1,r)e^{-\mu_1 r})' = -f(r)e^{-\mu_1 r}(p(x-1,1,0) + p(x-1,2,0)), \text{ for } x \geq 2.$$

Integrating this equation for $r$ from 0 to infinity leads to (28). Equation (27) can be deduced by removing the term $p(x-1,1,0)$ from (28).

Relations (25)-(28) allow us to relate the stationary probabilities $\pi_{x,1}$ and $\pi_{x,2}$, for $x \geq 1$. Since $p(x-1,2,0) + p(x-1,1,0) = \mu_2 \pi_{x,2} + \mu_1 \pi_{x,1}$, for $x \geq 2$ (Equation (23)) and $p(x,1,0) = \mu_2 \pi_{x,2}$, for $x \geq 1$ (Equation (24)), (28) leads to

$$\pi_{x,2} = \frac{\mu_1}{\mu_2} \frac{F^*(\mu_1)}{1 - F^*(\mu_1)} \pi_{x,1}, \text{ for } x \geq 2. \tag{29}$$

This relation is also valid for $x = 1$. This can be shown by combining (19), (24) for $x = 1$, and (27).

We now compute the LST of the probabilities $p(x,i,r)$. We multiply (2)-(5) by $e^{-sr}$ and integrate for $r$ from 0 to infinity. Recall that for a given function $h(r)$ with LST $h^*(s)$, we have

$\int_{r=0}^{\infty} e^{-sr} h'(r)\,\mathrm{d}r = sh^*(s) - h(0)$. This leads to

$$sP^*(0,2,s) - p(0,2,0) = -\mu_2 P^*(1,2,s) - \mu_1 P^*(1,1,s), \tag{30}$$

$$sP^*(x,2,s) - p(x,2,0) = \mu_2 P^*(x,2,s) - \mu_2 P^*(x+1,2,s) - \mu_1 P^*(x+1,1,s), \text{ for } x \geq 1, \tag{31}$$

$$sP^*(1,1,s) - p(1,1,0) = \mu_1 P^*(1,1,s) - F^*(s)p(0,2,0), \text{ and} \tag{32}$$

$$sP^*(x,1,s) - p(x,1,0) = \mu_1 P^*(x,1,s) - F^*(s)(p(x-1,2,0) + p(x-1,1,0)), \text{ for } x \geq 2. \tag{33}$$

Combining (32) and (33) with (23) and (29) leads to

$$P^*(x,1,s) = -\frac{\mu_2 \pi_{x,2}}{F^*(\mu_1)} \frac{F^*(s) - F^*(\mu_1)}{s - \mu_1}, \tag{34}$$

for $x \geq 1$. We now focus on $P^*(x,2,s)$ and $\pi_{x,2}$. We have for $x \geq 1$

$$\begin{aligned} p(x,2,0) &= \mu_2 \pi_{x+1,2} + \mu_1 \pi_{x+1,1} - p(x,1,0) \quad \text{(from (23))} \\ &= \mu_2 \pi_{x+1,2} + \mu_2 \frac{1 - F^*(\mu_1)}{F^*(\mu_1)} \pi_{x+1,2} - \mu_2 \pi_{x,2} \quad \text{(from (29))} \\ &= \mu_2 \left( \frac{1}{F^*(\mu_1)} \pi_{x+1,2} - \pi_{x,2} \right). \end{aligned}$$

Moreover, for $x \geq 1$, we have

$$\begin{aligned} p(x,2,0) &= \mu_2 P^*(x+1,2,\mu_2) + \mu_1 P^*(x+1,1,\mu_2) \quad \text{(from (26))} \\ &= \mu_2 P^*(x+1,2,\mu_2) - \frac{\mu_1 \mu_2 \pi_{x+1,2}}{F^*(\mu_1)} \frac{F^*(\mu_2) - F^*(\mu_1)}{\mu_2 - \mu_1} \quad \text{(from (34))}. \end{aligned}$$

This leads to

$$\pi_{x,2} = \frac{\pi_{x+1,2}}{F^*(\mu_1)} \left( 1 + \frac{\mu_1 (F^*(\mu_2) - F^*(\mu_1))}{\mu_2 - \mu_1} \right) - P^*(x+1,2,\mu_2), \text{ for } x \geq 1.$$

We also have for $x \geq 1$,

$$\begin{aligned} (s - \mu_2) P^*(x,2,s) &= p(x,2,0) - \mu_2 P^*(x+1,2,s) - \mu_1 P^*(x+1,1,s) \quad \text{(from (31))} \\ &= \mu_2 (P^*(x+1,2,\mu_2) - P^*(x+1,2,s)) + \mu_1 (P^*(x+1,1,\mu_2) - P^*(x+1,1,s)) \quad \text{(from (26))}. \end{aligned}$$

Finally, using (34) results in

$$P^*(x,2,s) = -\frac{\mu_1\mu_2}{F^*(\mu_1)(s-\mu_2)}\left(\frac{F^*(\mu_2)-F^*(\mu_1)}{\mu_2-\mu_1} - \frac{F^*(s)-F^*(\mu_1)}{s-\mu_1}\right)\pi_{x+1,2} - \mu_2\frac{P^*(x+1,2,s)-P^*(x+1,2,\mu_2)}{s-\mu_2}.$$

In summary, we may write for $x \geq 1$,

$$\pi_{x,2} = (g(0)+1)\pi_{x+1,2} - P^*(x+1,2,\mu_2), \text{ and} \tag{35}$$

$$P^*(x,2,s) = g(s)\pi_{x+1,2} - \mu_2\frac{P^*(x+1,2,s)-P^*(x+1,2,\mu_2)}{s-\mu_2}, \tag{36}$$

where $g(s) = -\frac{\mu_1\mu_2}{F^*(\mu_1)(s-\mu_2)}\left(\frac{F^*(\mu_2)-F^*(\mu_1)}{\mu_2-\mu_1} - \frac{F^*(s)-F^*(\mu_1)}{s-\mu_1}\right)$.

From (35), we show by induction that

$$\pi_{x,2} = -\sum_{k=0}^{\infty}(1+g(0))^k P^*(x+1+k,2,\mu_2), \tag{37}$$

for $x \geq 1$. From (36), we obtain by induction that

$$P^*(x,2,s) = \sum_{k=0}^{\infty}(-\mu_2)^k\frac{g(s)-T_{k-1}(g(s))}{(s-\mu_2)^k}\pi_{x+1+k,2}, \text{ and} \tag{38}$$

$$P^*(x,2,\mu_2) = \sum_{k=0}^{\infty}(-\mu_2)^k\frac{g^{(k)}(\mu_2)}{k!}\pi_{x+1+k,2}, \tag{39}$$

with $T_k(g(s)) = \sum_{j=0}^{k} g^{(j)}(\mu_2)\frac{(s-\mu_2)^j}{j!}$. Combining (37) and (39) allows us to show that the steady state probabilities have a similar relation as for a G/M/1 queue (for instance, see Kleinrock (1975), page 246). Therefore, the stationary probabilities have a geometric form, where $\pi_{x,2}$ can be written as $\pi_{x,2} = q^{x-1}\pi_{1,2}$, for $x \geq 1$. We thus deduce that

$$P^*(x,2,\mu_2) = q^x\pi_{1,2}\sum_{k=0}^{\infty}(-\mu_2 q)^k\frac{g^{(k)}(\mu_2)}{k!} = q^x\pi_{1,2}g(\mu_2(1-q)),$$

for $x \geq 1$. This leads to

$$\pi_{x,2} = -\pi_{1,2}\frac{q^{x+1}g(\mu_2(1-q))}{1-(1+g(0))q},$$

for $x \geq 1$. Therefore, $q$, is solution of

$$q^2 g(\mu_2(1-q)) = (1+g(0))q - 1. \tag{40}$$

This equation is equivalent to (9).

There remains to show that (9) has a unique solution in $(0,1)$. To this end, we consider the quantity $K$ define as

$$K = \frac{(\mu_1 - \mu_2)(1-q)F^*(\mu_1) + q\mu_1 F^*(\mu_2(1-q))}{\mu_1 - \mu_2(1-q)}.$$

We want to prove that the equation $K = q$ has a unique solution in $q$ with $q$ in $(0,1)$. First, we note that for $q = 0$, $K = F^*(\mu_1) > 0$ and for $q = 1$, $K = 1$. To show that there exists a unique solution of $K = q$ in $q$ for $q$ in $(0,1)$, we show that $K$ is increasing and convex in $q$. We may write

$$\frac{\partial K}{\partial q} = \int_{t=0}^{\infty} f(t)e^{-\mu_1 t} \frac{\mu_1 z(e^{zt}-1) + q\mu_1\mu_2(1 - e^{zt}(1-zt))}{z^2} \mathrm{d}t,$$

where $z = \mu_1 - \mu_2(1-q)$. We have $z(e^{zt}-1) \geq 0$, for $z \in \mathbb{R}$. Consider the function in $x$, $m(x) = 1 - e^x(1-x))$. We have $m'(x) = xe^x$. This function is positive for $x \geq 0$ and negative for $x \leq 0$. Therefore, $m(x)$ has a minimum at $x = 0$. Since $m(0) = 0$, $m(x) \geq 0$ for $x \in \mathbb{R}$. This proves that $1 - e^{zt}(1-zt) \geq 0$ and $\frac{\partial K}{\partial q} \geq 0$. Thus, $K$ is increasing in $q$. We next prove that $K$ is convex in $q$. We may express $\frac{\partial^2 K}{\partial q^2}$ as

$$\frac{\partial^2 K}{\partial q^2} = 2\mu_1\mu_2 \int_{t=0}^{\infty} f(t)e^{-\mu_1 t} \frac{e^{zt}(zt-1)+1}{z^2} \mathrm{d}t + \mu_1\mu_2^2 \int_{t=0}^{\infty} f(t)e^{-\mu_1 t} \frac{2q\left(e^{zt}\left(\frac{(zt)^2}{2} - zt + 1\right) - 1\right)}{z^3} \mathrm{d}t,$$

where $z = \mu_1 - \mu_2(1-q)$. As proven for $\frac{\partial K}{\partial q}$, we have $(1 + e^{zt}(zt-1)) \geq 0$. There remains to proven that $n(x) = e^x\left(1 - x + \frac{x^2}{2}\right) - 1 \geq 0$ if $x \geq 0$ and $n(x) \leq 0$ if $x \leq 0$. We have $n'(x) = \frac{x^2}{2}e^x \geq 0$. Therefore, $n(x)$ is increasing in $x$. Moreover, $n(0) = 0$. This proves that $n(x)$ has the sign of $x$ and $\frac{\partial^2 K}{\partial q^2} \geq 0$. Therefore, $K$ is convex in $q$.

In summary, we related the stationary probabilities $\pi_{x,1}$ and $\pi_{x,2}$ with $\pi_{1,2}$ for $x \geq 1$. There remains to relate $\pi_{1,2}$ with $\pi_{0,2}$. For this purpose, we need to determine $P^*(0,2,s)$ as given by (30). In this expression we need to express $P^*(1,2,s)$, $P^*(1,1,s)$ and $p(0,2,0)$. From the above results,

we get

$$P^*(1,2,s) = q\pi_{1,2}\frac{g(s)(s-\mu_2)+q\mu_2 g(\mu_2(1-q))}{s-\mu_2(1-q)},$$
$$P^*(1,1,s) = -\frac{\mu_2\pi_{1,2}}{F^*(\mu_1)}\frac{F^*(s)-F^*(\mu_1)}{s-\mu_1}, \text{ and}$$
$$p(0,2,0) = \mu_2\pi_{1,2}+\mu_1\pi_{1,1} = \frac{\mu_2}{F^*(\mu_1)}\pi_{1,2}.$$

This leads to

$$P^*(0,2,s) = \mu_2\pi_{1,2}\left(\frac{1}{sF^*(\mu_1)} - q\frac{g(s)(s-\mu_2)+q\mu_2 g(\mu_2(1-q))}{s(s-\mu_2(1-q))} + \frac{\mu_1}{sF^*(\mu_1)}\frac{F^*(s)-F^*(\mu_1)}{s-\mu_1}\right) \tag{41}$$
$$= \mu_2\pi_{1,2}\frac{F^*(s)\mu_1(s-\mu_2)+s(\mu_2-\mu_1)F^*(\mu_1)+(s-\mu_1)(s-\mu_2)}{F^*(\mu_1)s(s-\mu_1)(s-\mu_2(1-q))}$$

Equation (41) allows us to compute $P^*(0,2,s)$ as a function of $\pi_{1,2}$. We next deduce $\pi_{0,2} = \lim_{s\longrightarrow 0} P^*(0,2,s)$. Using $F^*(s) = 1 - E(A)s + o(s)$ as $s$ is in the neighborhood of zero, where $E(A)$ is the expected inter-arrival time, we deduce that

$$\pi_{0,2} = \pi_{1,2}\frac{\mu_2(\mu_1 E(A)-1)+(\mu_2-\mu_1)F^*(\mu_1)}{F^*(\mu_1)\mu_1(1-q)}.$$

The normalizing condition leads to the expression of $\pi_{0,2}$. This finishes the proof of the theorem. □

# C Proof of Corollary 1

In what follows, we explain how the performance measures can be derived. The expected service rate $\mu$ can be computed as

$$\mu = \frac{\mu_1\sum_{x=1}^{\infty}\pi_{x,1}+\mu_2\sum_{x=1}^{\infty}\pi_{2,1}}{\sum_{x=1}^{\infty}\pi_{x,1}+\sum_{x=1}^{\infty}\pi_{x,2}} = \frac{\mu_1\mu_2}{\mu_1 F^*(\mu_1)+\mu_2(1-F^*(\mu_1))}. \tag{42}$$

Therefore, the probability of am empty system $\pi_{0,2}$ can be expressed as

$$\pi_{0,2} = 1 - \frac{\mu_1 F^*(\mu_1)+\mu_2(1-F^*(\mu_1))}{\mu_1\mu_2 E(A)} = 1 - \frac{1}{\mu E(A)}. \tag{43}$$

Finally, we determine the expected waiting time in the system given through $E(W) = E(A)\sum_{x=1}^{\infty}(x-1)(\pi_{x,1}+\pi_{x,2})$. This leads to

$$E(W) = \frac{q}{\mu(1-q)}. \tag{44}$$

# D Proof of Proposition 1

*Proof.* We have

$$\frac{\partial^2 U}{\partial \mu_2^2} = -\frac{pr}{E(A)}\frac{\partial^2 E(W)}{\partial \mu_2^2} + (e-b)\frac{\partial^2 \mu}{\partial \mu_2^2}.$$

Therefore, to prove that the utility is concave in $\mu_2$, we prove that the expected waiting time $E(W)$ is decreasing and convex in $\mu_2$ and that the average service rate $\mu$ is increasing and concave in $\mu_2$. We may write

$$\frac{\partial E(W)}{\partial \mu_2} = \frac{\mu\frac{\partial q}{\partial \mu_2} - q(1-q)\frac{\partial \mu}{\partial \mu_2}}{(\mu(1-q))^2}, \text{ and,}$$
$$\frac{\partial^2 E(W)}{\partial \mu_2^2} = \frac{\mu^2(1-q)\frac{\partial^2 q}{\partial \mu_2^2} + 2\mu^2\left(\frac{\partial q}{\partial \mu_2}\right)^2 + 2q(1-q)^2\left(\frac{\partial \mu}{\partial \mu_2}\right)^2 - q\mu(1-q)^2\frac{\partial^2 \mu}{\partial \mu_2^2} - 2\mu(1-q)\frac{\partial \mu}{\partial \mu_2}\frac{\partial q}{\partial \mu_2}}{(\mu(1-q))^3}.$$

Moreover using (10), we get

$$\frac{\partial \mu}{\partial \mu_2} = \frac{\mu_1^2 F^*(\mu_1)}{(\mu_1 F^*(\mu_1) + \mu_2(1-F^*(\mu_1)))^2} > 0, \text{ and, } \quad \frac{\partial^2 \mu}{\partial \mu_2^2} = -\frac{2\mu_1^2 F^*(\mu_1)(1-F^*(\mu_1))}{(\mu_1 F^*(\mu_1) + \mu_2(1-F^*(\mu_1)))^3} < 0.$$

Therefore, $\mu$ is increasing and concave in $\mu_2$. Consequently, a necessary condition for $E(W)$ to be decreasing and convex in $\mu_2$ is to have $q$ also decreasing and convex in $\mu_2$.

From the proof of Theorem 1, to prove that $q$ is decreasing and convex in $\mu_2$, we need to show that the quantity $K$, defined in the proof of Theorem 1, is decreasing and convex in $\mu_2$. We may write

$$\frac{\partial K}{\partial \mu_2} = -\frac{q\mu_1(1-q)\mu_1}{z^2}\int_0^\infty f(t)e^{-\mu_1 t}\left(zte^{zt} + 1 - e^{zt}(1-q)\right)\mathrm{d}t, \text{ and,}$$
$$\frac{\partial^2 K}{\partial \mu_2^2} = \frac{2q\mu_1(1-q)^2}{z^3}\int_0^\infty f(t)e^{-\mu_1 t}\left(-1 + \left(\frac{(zt)^2}{2} - zt + 1\right)e^{zt}\right)\mathrm{d}t,$$

with $z = \mu_1 - \mu_2(1-q)$. The functions involved in $\frac{\partial K}{\partial \mu_2}$ and $\frac{\partial^2 K}{\partial \mu_2^2}$ are those already studied in the proof of Theorem 1. In particular, we showed that $1 + (zt-1)e^{zt} \geq 0$ for $z \in \mathbb{R}$ and that $-1 + \left(\frac{(zt)^2}{2} - zt + 1\right) e^{zt} \geq 0$ if and only if $z \geq 0$. This proves that $\frac{\partial K}{\partial \mu_2} \leq 0$ and $\frac{\partial^2 K}{\partial \mu_2^2} \geq 0$. Consequently, this also proves that $E(W)$ is decreasing and convex in $\mu_2$.

For $\mu_2 \leq \mu_1$, the agent's utility can be expressed as $U = FW - e\mu_1 + \frac{pr}{E(A)}(E(W_e) - E(W)) + (e-b)\mu$. Since $E(W)$ is decreasing in $\mu_2$ and $\mu$ is increasing in $\mu_2$, $U$ is also increasing in $\mu_2$ for $\mu_2 \leq \mu_1$. This proves that the optimal value of $\mu_2$ should be selected such that $\mu_2 \geq \mu_1$. The optimal service rate $\mu_2$ is solution of $\frac{\partial U}{\partial \mu_2} = 0$, for $\mu_2 \geq \mu_1$, where

$$\frac{\partial U}{\partial \mu_2} = -\frac{pr}{E(A)} \frac{\partial E(W)}{\partial \mu_2} - e + (e-b)\frac{\partial \mu}{\partial \mu_2}.$$

Thus, the optimal piece-rate compensation is given by

$$pr = \frac{\left(e - (e-b)\frac{\partial \mu}{\partial \mu_2}\right) E(A)}{-\frac{\partial E(W)}{\partial \mu_2}}.$$

Since $\frac{\partial E(W)}{\partial \mu_2} < 0$, we only need to prove that $e-(e-b)\frac{\partial \mu}{\partial \mu_2} \geq 0$ to prove that $pr \geq 0$. The inequality $e-(e-b)\frac{\partial \mu}{\partial \mu_2} \geq 0$ is equivalent to $\frac{\mu_1^2 F^*(\mu_1)}{(\mu_1 F^*(\mu_1) + \mu_2(1-F^*(\mu_1)))^2} \leq \frac{e}{e-b}$. Since $b \leq e$, we have $\frac{e}{e-b} \geq 1$. To show that the inequality $e-(e-b)\frac{\partial \mu}{\partial \mu_2} \geq 0$ holds, we now prove that $\frac{\mu_1^2 F^*(\mu_1)}{(\mu_1 F^*(\mu_1) + \mu_2(1-F^*(\mu_1)))^2} \leq 1$. This last inequality is equivalent to $(1-F^*(\mu_1))(\mu_2^2(1-F^*(\mu_1)) + 2\mu_1\mu_2 F^*(\mu_1) - \mu_1^2 F^*(\mu_1)) \geq 0$. We have $F^*(\mu_1) \leq 1$. The function $\mu_2^2(1-F^*(\mu_1)) + 2\mu_1\mu_2 F^*(\mu_1) - \mu_1^2 F^*(\mu_1)$ is increasing in $\mu_2$. Since $\mu_2 \geq \mu_1$, we have $\mu_2^2(1-F^*(\mu_1)) + 2\mu_1\mu_2 F^*(\mu_1) - \mu_1^2 F^*(\mu_1) \geq \mu_1^2(1-F^*(\mu_1)) + 2\mu_1^2 F^*(\mu_1) - \mu_1^2 F^*(\mu_1) = \mu_1^2 \geq 0$. This proves that $(1-F^*(\mu_1))(\mu_2^2(1-F^*(\mu_1)) + 2\mu_1\mu_2 F^*(\mu_1) - \mu_1^2 F^*(\mu_1)) \geq 0$. Consequently, this proves that $pr \geq 0$.

Next, we have

$$\frac{\partial pr}{\partial \mu_2} = E(A) \frac{(e-b)\frac{\partial^2 \mu}{\partial \mu_2^2}\frac{\partial E(W)}{\partial \mu_2} + \left(e - (e-b)\frac{\partial \mu}{\partial \mu_2}\right)\frac{\partial^2 E(W)}{\partial \mu_2^2}}{\left(\frac{\partial E(W)}{\partial \mu_2}\right)^2}.$$

To show that $pr \geq 0$, we already proved that $e-(e-b)\frac{\partial \mu}{\partial \mu_2} \geq 0$. Moreover, we also proved that

$\frac{\partial^2 \mu}{\partial \mu_2^2} \leq 0$, $\frac{\partial E(W)}{\partial \mu_2} \leq 0$ and $\frac{\partial^2 E(W)}{\partial \mu_2^2} \geq 0$. Therefore, we have $\frac{\partial pr}{\partial \mu_2} \geq 0$. Finally, we may write

$$\frac{\partial C}{\partial \mu_2} = b\frac{\partial \mu}{\partial \mu_2} + \frac{(e-b)(E(W_e)-E(W))\frac{\partial^2 \mu}{\partial \mu_2^2}\frac{\partial E(W)}{\partial \mu_2} - \left(e-(e-b)\frac{\partial \mu}{\partial \mu_2}\right)\frac{\partial E(W)}{\partial \mu_2}\frac{\partial^2 E(W)}{\partial \mu_2^2}}{\left(-\frac{\partial E(W)}{\partial \mu_2}\right)^2}.$$

Using the monotonicity results in $\mu_2$ for $\mu$ and $E(W)$, we deduce from this expression that $\frac{\partial C}{\partial \mu_2} > 0$. Note that since $\frac{\partial \mu}{\partial \mu_2} > 0$, we also have $\frac{\partial r}{\partial \mu} \geq 0$ and $\frac{\partial C}{\partial \mu} \geq 0$. □

# E Proof of Lemma 2

*Proof.* By computing the derivative of (10) in $\mu_1$ assuming that $\mu$ is kept constant, we obtain after simplification

$$\frac{\partial \mu_2}{\partial \mu_1} = -\mu_2 \frac{\mu_2(1-F(\mu_1)) + \mu_1(\mu_2-\mu_1)F'(\mu_1)}{\mu_1^2 F(\mu_1)}.$$

In the case $\mu_1 \geq \mu_2$, we have $\frac{\partial \mu_2}{\partial \mu_1} \leq 0$, since $F'(\mu_1) < 0$ and $F(\mu_1) \leq 1$. We now consider the case $\mu_2 > \mu_1$. We may write

$$\mu_2(1-F(\mu_1)) + \mu_1(\mu_2-\mu_1)F'(\mu_1) = \int_{t=0}^{\infty} f(t)\left[\mu_2 - \mu_2 e^{-\mu_1 t} - \mu_1 t(\mu_2-\mu_1)e^{-\mu_1 t}\right] \mathrm{d}t.$$

We define the function in $t$, $w(t) = \mu_2 - \mu_2 e^{-\mu_1 t} - \mu_1 t(\mu_2-\mu_1)e^{-\mu_1 t}$. We obtain $w'(t) = \mu_1^2 e^{-\mu_1 t}\left(1+(\mu_2-\mu_1)t\right) > 0$, since $\mu_2 > \mu_1$. Therefore, $w(t)$ is increasing in $t$. Finally, $w(0) = 0$. This proves that $w(t) \geq 0$ and consequently $\mu_2(1-F(\mu_1)) + \mu_1(\mu_2-\mu_1)F'(\mu_1) \geq 0$. Hence, also in this case we have $\frac{\partial \mu_2}{\partial \mu_1} \leq 0$. □

# F Proof of Proposition 2

*Proof.* We employ the notation $f(x) \underset{x \to a}{\sim} g(x)$ to indicate that $\lim_{x \to a} \frac{f(x)}{g(x)} = 1$. From the theorem of the initial value, we have $\lim_{\mu_1 \to \infty} \mu_1 F^*(\mu_1) = f(0)$. Therefore, from (14), we have $\mu_2 \underset{\mu_1 \to \infty}{\sim} \frac{\mu f(0)}{\mu_1}$, if

$f(0) \neq 0$. Therefore, we have

$$\begin{aligned}
&- q\mu_1 F^*(\mu_2(1-q)) + (\mu_2-\mu_1)(1-q)F^*(\mu_1) - (\mu_2(1-q)-\mu_1)q \\
&\underset{\mu_1\to\infty}{\sim} q\mu_1\left(1 - F^*\left(\frac{\mu f(0)(1-q)}{\mu_1}\right)\right) - f(0)(1-q) + \frac{\mu f(0)}{\mu_1}(1-q)(F^*(\mu_1)-q) \\
&\underset{\mu_1\to\infty}{\sim} qE(A)\mu f(0)(1-q) - f(0)(1-q).
\end{aligned}$$

Therefore, we should have $q = \frac{1}{\mu E(A)}$ to solve (40) as $\mu_1$ tends to infinity. This proves the first statement in the case $f(0) \neq 0$. The same result can be proven in the case $f(0) = 0$ by considering a sequence of density functions, $f_n(t)$, which tends uniformly to $f(t)$ as $n$ tends to infinity, with $f_n(0) \neq 0$.

We now consider the second statement.

$$\begin{aligned}
&- q\mu_1 F^*(\mu_2(1-q)) + (\mu_2-\mu_1)(1-q)F^*(\mu_1) - (\mu_2(1-q)-\mu_1)q \\
&\underset{\mu_2\to\infty}{\sim} \mu_2(F^*(\mu_1)-q)(1-q) + \mu_1(1-q)\left(\frac{q}{1-q} - F^*(\mu_1)\right).
\end{aligned}$$

Therefore, the last expression tends to zero if $q$ can be expressed as $q = F^*(\mu_1) + \frac{a}{\mu_2} + o\left(\frac{1}{\mu_2}\right)$, when $\mu_2$ tends to infinity. When $\mu_2$ tends to infinity, we should also have $\mu_1 = \mu(1 - F^*(\mu_1))$, from (14). Combining $q = F^*(\mu_1)$ and $\mu_1 = \mu(1 - F^*(\mu_1))$ leads to $q = F^*(\mu(1-q))$ which corresponds to the equation for $q$ in the case $\mu_1 = \mu_2$. Therefore, $q$ tends to $q_e$ when $m_2$ tends to infinity. □

# G Proof of Theorem 2

*Proof.* To simplify the notation in the proof, we write $F$ instead of $F^*$. As a first step, we prove that the expected wait has a unique local extremum in $\mu_1$ when $\mu_1$ and $\mu_2$ are related via (14). From (12), we have

$$\frac{\partial E(W)}{\partial \mu_1} = \frac{1}{\mu}\frac{\frac{\partial q}{\partial \mu_1}}{(1-q)^2}.$$

Therefore, $\frac{\partial E(W)}{\partial \mu_1} = 0$ if and only if $\frac{\partial q}{\partial \mu_1} = 0$. Next using (9), we deduce that

$$\begin{aligned}
&\frac{\mu_1}{q}\frac{\partial q}{\partial \mu_1}\left((\mu_1-\mu_2)F(\mu_1)+q^2\mu_2+q^2\mu_1\mu_2F'(\mu_2(1-q))\right) \qquad (45)\\
&\quad=\frac{\partial \mu_2}{\partial \mu_1}\mu_1(1-q)\left(q-F(\mu_1)+q\mu_1F'(\mu_2(1-q))\right)+(1-q)\left((\mu_1-\mu_2)\mu_1F'(\mu_1)-\mu_2q+\mu_2(1-q)F(\mu_1)\right).
\end{aligned}$$

Therefore, if $\frac{\partial q}{\partial \mu_1} = 0$, we should have

$$\psi(q)=\frac{\partial \mu_2}{\partial \mu_1}\mu_1\left(q-F(\mu_1)+q\mu_1F'(\mu_2(1-q))\right)+\left((\mu_1-\mu_2)\mu_1F'(\mu_1)-\mu_2q+\mu_2(1-q)F(\mu_1)\right)=0, \qquad (46)$$

since $q \neq 1$. Therefore, under Contract 2, a critical point is determined by $(\mu_1, \mu_2, q)$ such that Equations (14), (9), and (46) are satisfied by $(\mu_1, \mu_2, q)$. Since these three equations are independent, there is at most one critical point which could either be a maximum or a minimum of $E(W)$.

Using (14), we rewrite the equation leading to $q$ as a function of $\mu_1$ and $\mu$ as follows:

$$\begin{aligned}
&-q\mu_1F\left(\frac{\mu\mu_1F(\mu_1)(1-q)}{\mu_1-\mu(1-F(\mu_1))}\right)+\left(\frac{\mu\mu_1F(\mu_1)}{\mu_1-\mu(1-F(\mu_1))}-\mu_1\right)(1-q)F(\mu_1)\\
&\quad-\left(\frac{\mu\mu_1F(\mu_1)(1-q)}{\mu_1-\mu(1-F(\mu_1))}-\mu_1\right)q=0.
\end{aligned}$$

Multiplying this equation by $\frac{\mu_1-\mu(1-F(\mu_1))}{\mu_1}$ leads to

$$-q(\mu_1-\mu(1-F(\mu_1))F\left(\frac{\mu\mu_1F(\mu_1)(1-q)}{\mu_1-\mu(1-F(\mu_1))}\right)+(\mu-\mu_1)(1-q)F(\mu_1)-q(\mu-\mu_1-q\mu F(\mu_1))=0.$$

We consider a couple $(\mu_1, \mu_2)$ which is close to the point $(\mu_1 = \mu, \mu_2 = \mu)$. We then write $\mu_1 = \mu+\epsilon$, where $\epsilon = o(\mu)$. Therefore, we have

$$-q(\epsilon+\mu F(\mu+\epsilon))F\left(\frac{\mu(1-q)(\mu+\epsilon)F(\mu+\epsilon)}{\epsilon+\mu F(\mu+\epsilon)}\right)-\epsilon(1-q)F(\mu+\epsilon)+q(q\mu F(\mu+\epsilon)+\epsilon)=0. \qquad (47)$$

We have

$$
\begin{aligned}
&F\left(\frac{\mu(1-q)(\mu+\epsilon)F(\mu+\epsilon)}{\epsilon+\mu F(\mu+\epsilon)}\right) = F\left((1-q)\mu\frac{\mu F(\mu)+\epsilon(F(\mu)+\mu F'(\mu))+o(\epsilon)}{\mu F(\mu)+\epsilon(1+\mu F'(\mu))+o(\epsilon)}\right)\\
&= F\left((1-q)\mu\left(1+\frac{\epsilon}{\mu}\left(1+\frac{F'(\mu)}{F(\mu)}\right)+o(\epsilon)\right)\left(1-\frac{\epsilon}{\mu}\left(\frac{1}{F(\mu)}+\frac{F'(\mu)}{F(\mu)}\right)+o(\epsilon)\right)\right)\\
&= F\left((1-q)\mu\left(1+\frac{\epsilon}{\mu}\left(1-\frac{1}{F(\mu)}\right)+o(\epsilon)\right)\right)\\
&= F((1-q)\mu)+\epsilon(1-q)\left(1-\frac{1}{F(\mu)}\right)F'(\mu(1-q))+o(\epsilon).
\end{aligned}
$$

Moreover, $(\epsilon+\mu F(\mu+\epsilon)) = \mu F(\mu)+\epsilon(1+\mu F'(\mu))+o(\epsilon)$. Hence,

$$
\begin{aligned}
&-q(\epsilon+\mu F(\mu+\epsilon))F\left(\frac{\mu(1-q)(\mu+\epsilon)F(\mu+\epsilon)}{\epsilon+\mu F(\mu+\epsilon)}\right)\\
&= -q\mu F(\mu)F((1-q)\mu)+\epsilon\left[-qF((1-q)\mu)(1+\mu F'(\mu))+q(1-F(\mu))(1-q)\mu F'(\mu(1-q))\right]+o(\epsilon).
\end{aligned}
$$

Next,

$$
-\epsilon(1-q)F(\mu+\epsilon)+q(q\mu F(\mu+\epsilon)+\epsilon) = q^2\mu F(\mu)+\epsilon(q+q^2\mu F'(\mu)-(1-q)F(\mu))+o(\epsilon).
$$

Therefore, (47) can be rewritten as

$$
\begin{aligned}
&-q\mu F(\mu)F((1-q)\mu)+q^2\mu F(\mu)\\
&+\epsilon\left[-qF((1-q)\mu)(1+\mu F'(\mu))+q(1-F(\mu))(1-q)\mu F'(\mu(1-q))+q+q^2\mu F'(\mu)-(1-q)F(\mu)\right]+o(\epsilon)=0.
\end{aligned}
$$

We now write $q = q_e + z$, where $q_e = F(\mu(1-q_e))$ (that is, the solution when $\mu_1=\mu_2=\mu$), and $z = o(q_e)$. Using $q_e = F(\mu(1-q_e))$, we obtain

$$
q_e z\mu F(\mu)(1+\mu F'((1-q_e)\mu)) \underset{\epsilon,z\to 0}{\sim} -\epsilon(1-q_e)(q_e-F(\mu)+q_e\mu(1-F(\mu))F'(\mu(1-q_e))). \quad (48)
$$

The equivalence (48) relates $z$ and $\epsilon$. On the left hand side of this expression the terms $q_e$ and $\mu F(\mu)$ are positive. Recall that $q_e$ is solution of $q = F(\mu(1-q))$, with $0<q<1$. Consider the function in $q$, $H(q) = F(\mu(1-q))$. We have $H'(q) = -\mu F'(\mu(1-q)) = \mu\int_{t=0}^{\infty} te^{-\mu(1-q)t}f(t)\mathrm{d}t > 0$, and $H''(q) = \mu^2 F''(\mu(1-q)) = \mu^2\int_{t=0}^{\infty} t^2 e^{-\mu(1-q)t}f(t)\mathrm{d}t > 0$. Therefore, $H(q)$ is increasing and convex in $q$. Moreover, we have $H(1) = F(0) = 1$, so $q=1$ is a solution of $H(q)=q$, and $H(0) = F(\mu) > 0$.

If $H'(q_e) > 1$, then for $q > q_e$ we have $H'(q) > H'(q_e) > 1$, since $H(q)$ is convex in $q$. Therefore, we should have $H(1) = 1 > H'(q_e)(1 - q_e) + q_e > 1$ which leads to a contradiction. Therefore, we have $H'(q_e) \leq 1$. Hence, $-\mu F'(\mu(1 - q_e)) \leq 1$ and $1 - \mu F'(\mu(1 - q_e)) \geq 0$. This means that the terms proportional with $z$ are all positive.

With unequal rates, we want to find a solution with $q < q_e$ (as the expected waiting time is increasing in $q$). Therefore, we want to have $z < 0$. Therefore, if $\phi(q_e, \mu) > 0$, then we need to have $\epsilon > 0$ to have $z < 0$ and if $\phi(q_e, \mu) < 0$, then we need to have $\epsilon < 0$ to have $z < 0$.

Consequently if $\phi(q_e, \mu) < 0$, then there exists a couple $(\mu_1, \mu_2)$ with $\mu_1 < \mu_2$ such that the expected wait is lower than in the case with equal service rates. From Proposition 2, the $G/M^{event}/1$ queue behaves as in the case with equal service rates when $\mu_2$ tends to infinity. Therefore, there exists a minimum of $E(W)$ in $\mu_2$, for $\mu < \mu_2 < \infty$. Since $E(W)$ has at most one critical point in $\mu_2$, this critical point is the global minimum of $E(W)$ in $\mu_2$.

In the case $\phi(q_e, \mu) > 0$, there exists a couple $(\mu_1, \mu_2)$ with $\mu_1 > \mu_2$ such that the expected wait is lower than in the case with equal service rates. From Proposition 2, the $G/M^{event}/1$ queue behaves as an M/M/1 queue with service rate $\mu$ when $\mu_1$ tends to infinity. If the variability of the inter-arrival time is lower than the one of an exponential distribution, then there exists a minimum of $E(W)$ in $\mu_1$, for $\mu < \mu_1 < \infty$. Since $E(W)$ has at most one critical point in $\mu_1$, this critical point is the global minimum of $E(W)$ in $\mu_1$. However, the result holds even if the variability of the inter-arrival time is higher than the one of an exponential distribution. To show this result, we prove that $\mu_1 = \infty$ cannot be a minimum for $q$ and consequently for $E(W)$. To prove this result we consider Equation (45). Using the initial value theorem as we did in the proof of Proposition 2 and assuming that $f(0) \neq 0$, the left hand side of (45) is equivalent to $\mu_1 f(0)(\mu E(A) - 1)\frac{\partial q}{\partial \mu_1}$ when $\mu_1$ tends to infinity. Assuming that $tf'(t)$ tends to 0 as $t$ tends to 0, using again the initial value theorem, we show that $(1 - q)\left((\mu_1 - \mu_2)\mu_1 F'(\mu_1) - \mu_2 q + \mu_2(1 - q)F(\mu_1)\right)$ is equivalent to $-f(0)\left(1 + \frac{1}{\mu_1 E(A)}\right)\left(1 - \frac{1}{\mu E(A)}\right)$. Recall that we have

$$\frac{\partial \mu_2}{\partial \mu_1} = -\mu_2 \frac{\mu_2(1 - F(\mu_1)) + \mu_1(\mu_2 - \mu_1)F'(\mu_1)}{\mu_1^2 F(\mu_1)}.$$

This expression is equivalent to $\frac{-f(0)\mu}{\mu_1^2}\left(1 + \frac{\mu}{\mu_1}\right)$ as $\mu_1$ tends to infinity. Finally, $\mu_1(1 - q)\left(q - F(\mu_1) + q\mu_1 F'(\mu_2(1 - q))\right)$ is equivalent to $-\left(\frac{\mu_1^2}{\mu} + \frac{\mu_1}{\mu E(A)}\right)\left(1 - \frac{1}{\mu E(A)}\right)$. We then

deduce that

$$\frac{\partial q}{\partial \mu_1} \underset{\mu_1 \to \infty}{\sim} \frac{1}{\mu_1^2 E(A)} > 0.$$

This proves that $q$ cannot be decreasing in $\mu_1$ when $\mu_1$ is in the neighborhood of infinity. Therefore, the condition $\phi(q_e, \mu) > 0$ is sufficient to prove that $E(W)$ is minimized for a couple $(\mu_1, \mu_2)$ with $\mu_1 > \mu_2$. □